

Lehrinhalte von



Wegweiser.UX-für-KI

Unser Projekt bietet Hilfestellungen und praktische Informationen über **Künstliche Intelligenz (KI) und User Experience (UX) von KI-Systemen**, die auf das **Gemeinwohl** ausgerichtet sind.

Die Inhalte wurden für **Entwickler:innen, Projektmanager:innen und KI- oder UX-Interessierte** unabhängig von ihrem jeweiligen Erfahrungsniveau erstellt.

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

Was erwartet Sie?

In den folgenden Modulen präsentieren wir mithilfe von Videoeinheiten und Texten Inhalte zu relevanten Themen in den Bereichen UX und der Anwendung von KI-Systemen.



<https://youtu.be/LDvVRvG7OB0>

Wie nutzen Sie die Plattform optimal?

Die einzelnen Lektionen können **unabhängig voneinander** konsumiert werden. Dieser Aufbau erlaubt es Ihnen, für Sie interessante Themen auszuwählen und sich die Inhalte dazu anzusehen, oder aber der von uns erdachten Struktur zu folgen. Beachten Sie bitte, dass einzelne Lehrinhalte das Wissen aus vorherigen Modulen voraussetzen können. Hatten Sie zum Beispiel bisher wenig Berührungs punkte mit UX, kann es sinnvoll sein, die entsprechende Lehreinheit zu sehen, bevor Sie sich mit UX bezogenen KI-Eigenschaften auseinandersetzen.

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

Modul:

01 UX und Usability

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

01

Einleitung

[Kursübersicht](#) > [UX und Usability](#)

In diesem Modul gibt Ihnen Prof. Dr. Hans-Christian Jetter einen Überblick über die Begriffe **Usability** und **User Experience** – zwei zentrale Aspekte menschzentrierter Designprozesse, die in Forschung und Entwicklung eine wichtige Rolle spielen. Der Fokus auf beide Aspekte während der Entwicklung eines Produkts kann einen starken Einfluss auf das Gesamterlebnis von Nutzenden bei der Interaktion haben. Wir erklären Ihnen, wieso ein fehlender Fokus auf diese Aspekte ein Produkt unattraktiver für Kunden und Nutzende machen kann, und welche Punkte Sie beim Design Ihres Produkts beachten sollten. Abschließend erhalten Sie Informationen und Tipps dazu, wie die beiden Aspekte evaluiert werden können.

Im folgenden Video wird ein Überblick über die fünf wichtigen Begriffe des Moduls gegeben und warum sie wichtig sind.



<https://youtu.be/9cgmeenl72s>

Wichtige Begriffe

1

Usability

Beschreibt, wie effektiv, effizient und zufriedenstellend Nutzende ihre Ziele mit einem System erreichen. Eine gute Usability ist entscheidend für Akzeptanz und Erfolg eines Produkts.

2

Nutzungskontext

Bezieht sich auf die Rahmenbedingungen, unter denen ein System genutzt wird, einschließlich Nutzendenprofil, Aufgaben, Umgebung und Hilfsmittel. Ein tiefes Verständnis des Nutzungskontexts ist essenziell für eine benutzerzentrierte Gestaltung.

3

User Experience

Umfasst das gesamte Nutzungserlebnis eines Produkts, einschließlich subjektiver Empfindungen wie Ästhetik, Vertrauen und Freude. UX geht über Usability hinaus und ist heute ein Schlüssel zum Produkterfolg.

4

Mensch-Computer-Interaktion

Eine interdisziplinäre Forschungsdisziplin, die sich mit der Gestaltung von Interaktionen zwischen Menschen und Technologie beschäftigt. Ziel ist es, positive Nutzungserfahrungen durch optimierte Design- und Entwicklungsprozesse zu schaffen.

5

Evaluation

Bezeichnet die Überprüfung und Messung von Usability und UX eines Systems. Durch Methoden wie Tests und Nutzerfeedback wird ermittelt, ob die gewünschten Nutzungserfahrungen erreicht werden.

02 Usability

[Kursübersicht](#) > [UX und Usability](#)

Beschreibt, wie effektiv, effizient und zufriedenstellend Nutzende ihre Ziele mit einem System erreichen. Eine gute Usability ist entscheidend für Akzeptanz und Erfolg eines Produkts.

Definition Usability

Usability, im Fachjargon auch als „Gebrauchstauglichkeit“ bezeichnet, beschreibt die Qualität der Interaktion zwischen Menschen und interaktiven Systemen. In Alltagssprache würden wir am ehesten von „Benutzerfreundlichkeit“ oder von einer „intuitiven Bedienung“ sprechen.

Gute Usability ist eine Grundvoraussetzung für erfolgreiche Mensch-Computer-Interaktion. Undurchdachte oder schlecht gestaltete Systeme führen dagegen schnell zu Bedienproblemen, kognitiver Überlastung und Frustration bei den Nutzer:innen - also zu einer schlechten Usability.

Schon seit den 1980er Jahren wissen wir, dass gute Usability über den Erfolg und die Akzeptanz eines neuen Systems oder Produkts entscheidet. Digitale Systeme müssen schlicht und einfach eine gute

Usability haben, um überhaupt in der Praxis akzeptiert und verwendet werden zu können.

Drei Faktoren der Usability

Um Usability gezielter gestalten und messen zu können, definiert die Industrienorm DIN EN ISO 9241-11 drei Faktoren der Usability:

Effektivität, Effizienz und Zufriedenstellung.

1

Effektivität bedeutet, dass Nutzer:innen mit einem System ihre Aufgaben erfolgreich bewältigen können.

2

Effizienz beschreibt, dass dies mit einem angemessenen Aufwand geschieht.

3

Zufriedenstellung bezieht sich darauf, dass die Verwendung des Systems für ihre Aufgaben bei den Nutzer:innen keine starken negativen Emotionen wie Frustration oder Ärger auslöst.

In unserem alltäglichen Sprachgebrauch differenzieren wir dabei typischerweise nicht zwischen Effektivität und Effizienz. Dies ist aber sinnvoll, wenn wir von der Usability eines interaktiven Systems sprechen.

Unterschied: Effektivität und Effizienz

Ein einfaches Beispiel macht den Unterschied klar: Stellen Sie sich vor, Sie verwenden ein interaktives System, um einen dreiseitigen Brief an eine Behörde zu schreiben und als PDF hochzuladen.

Fall 1: Sie verwenden dazu einen PC oder Laptop mit einer gängigen Textverarbeitung.

Durch die Verwendung einer grafischen Oberfläche mit Textverarbeitungsprogramm, Maus und Tastatur werden Sie es mit großer Sicherheit schaffen, den Brief ohne Tippfehler, im erwünschten Layout und mit der gewünschten Länge zu verfassen. Damit ist das System für diese Aufgabe schon einmal grundsätzlich **effektiv**, denn Sie erreichen damit Ihr Ziel.

Mit etwas Übung sollte das Ganze auch innerhalb von 1-2 Stunden erledigt sein und der kognitive Aufwand ist für Sie überschaubar. Das setzt natürlich voraus, dass Sie der Sprache des Briefes mächtig sind, einigermaßen Übung im Schreiben und Tippen haben und mit dem Programm einigermaßen vertraut sind. Damit ist das System für diese Aufgabe **effizient**, da der Aufwand angemessen ist.

Es ist auch nicht zu erwarten, dass das System Sie dabei erheblich stressst oder ärgert. Es ist also im Großen und Ganzen **zufriedenstellend**.

Unterm Strich ist damit eine **gute Usability** erreicht.

Fall 2: Sie verwenden dazu ein Mobiltelefon und verfassen den Text über die Texteingabe auf dem Touchscreen in einer einfachen Notizen-App.

Auch hier werden Sie es letztendlich irgendwann schaffen, den Brief ohne Tippfehler und weitgehend im erwünschten Layout und in der erwünschten Länge zu schreiben. Das System ist also für Ihre Aufgabe grundsätzlich **effektiv**.

Allerdings wird dies aufgrund des kleinen Bildschirms, des umständlichen Einrückens von Text über Leerzeichen, der mangelnden Rechtschreibkorrektur, etc. wahrscheinlich ein sehr langer und kognitiv anstrengender Prozess. Das System ist also **nicht effizient**.

Man kann sogar davon ausgehen, dass Sie der Prozess belasten wird und es auch Phasen geben wird, in der Sie die Bearbeitung der Aufgabe mit dem System frustriert und ärgert. Somit entsteht dann auch eine **schlechte Zufriedenstellung**.

Unterm Strich ist es damit eine **schlechte Usability** - trotz prinzipieller Effektivität.

Fall 3: Sie haben bedingt durch Ihre Sehfähigkeit oder durch die Beweglichkeit ihrer Finger, Hände und Arme besondere Anforderungen im Bereich der Barrierefreiheit.

Beispielsweise könnten Sie vielleicht den Text des Briefes weder auf dem PC-Bildschirm noch dem Telefon-Bildschirm gut lesen. Oder die Texteingabe über eine physische PC-Tastatur oder eine Bildschirmtastatur auf dem Touchscreen ist für Sie faktisch unbedienbar. Für Sie persönlich sind die Systeme aus Fall 1 und 2 somit **nicht effektiv, nicht effizient und nicht zufriedenstellend** ist, wobei sie es für andere Personen sein mögen.

Fazit aus den Fallbeispielen

Es ist wichtig zu verstehen, dass ein System nicht „eine“ Usability hat, sondern dass Effektivität, Effizienz und Zufriedenstellung immer davon abhängig sind, wer ein System verwendet, zu welchem Zweck es verwendet wird und wie und wo damit interagiert wird. **Usability ist also keine Systemeigenschaft, sondern ergibt sich immer aus dem Kontext der Verwendung.**

Im Bezug auf Usability gibt es daher keine inhärent „guten“ oder „schlechten“ Systeme, sondern es hängt immer davon ab, für was und von wem etwas verwendet wird. Oder wie es der vielbeachtete Usability-Experte und UX-Designer Bill Buxton auf den Punkt bringt:

„Everything is best for something and worst for something else.“

Usability ist immer die Verwendbarkeit und Praxistauglichkeit eines Systems durch eine bestimmte Person und für eine bestimmte Aufgabe. Die genaue Definition in der DIN EN ISO 9241-11 bringt dies zum Ausdruck:

„Die Gebrauchstauglichkeit ist das Ausmaß, in dem ein System, ein Produkt oder eine Dienstleistung durch bestimmte Benutzer genutzt werden kann, um in einem bestimmten Nutzungskontext bestimmte Ziele effektiv, effizient und zufriedenstellend zu erreichen.“

Ausblick

Usability ist dabei eine notwendige Grundanforderung für den Erfolg eines Systems oder Produkts. Ganz entscheidend ist dafür, dass der reale **Nutzungskontext** schon frühzeitig bei der Konzeption, Gestaltung und Entwicklung des Systems berücksichtigt wird.

Was sich genau hinter diesem Begriff verbirgt, werden wir in der nächsten Lektion thematisieren.

03

Nutzungskontext

Kursübersicht > UX und Usability

Beschreibt den Rahmen, in dem ein System verwendet wird. Dazu gehören die Nutzer:innen, ihre Aufgaben, ihre Umgebung und ihre Hilfsmittel. Man muss den Nutzungskontext verstehen, um Systeme wirklich benutzerzentriert zu gestalten und einsetzen zu können.

Definition Nutzungskontext

Der Nutzungskontext beschreibt die Umstände, unter denen ein digitales System benutzt werden wird. Wer also ein benutzerfreundliches System gestalten will, sollte frühzeitig mit den Nutzer:innen zusammenarbeiten und folgende Aspekte des Nutzungskontext ermitteln:

1

die **Benutzer:innen**, z.B. ihre Kenntnisse, Erfahrungen und Erwartungen

2

die **Aufgaben**, die mit dem System erledigt werden sollen

3

die verwendeten **Arbeitsmittel**, z.B. Software, Geräte

4

die **physische und soziale Umgebung**, z.B. die Art, wie der Arbeitsplatz gestaltet ist oder die Menschen, mit denen man arbeitet

Warum ist der Nutzungskontext wichtig?

Wenn Sie die Usability (Gebrauchstauglichkeit) eines interaktiven Systems bewerten oder verbessern möchten, muss der jeweilige Nutzungskontext und dessen Anforderungen bekannt sein. Das System ist nicht automatisch gebrauchstauglich, weil es bestimmte Funktionen bietet. Es kommt darauf an, **wie gut das System in der Praxis funktioniert** (also in seinen realen Nutzungskontexten) und ob es von der Zielgruppe dabei **effektiv, effizient und zufriedenstellend** genutzt werden kann.



Nutzer:innen



Aufgaben



Produkte



Effektivität



Effizienz



Zufriedenheit

Usability

Nutzungskontext

Physische und soziale Umgebung inklusive Hardware,
Software und Materialien

Daher ist es wichtig, frühzeitig mit der Zielgruppe in Kontakt zu treten und ihre Perspektive bei der Konzeption und Entwicklung mit einzubeziehen. Das kann die spätere Akzeptanz und den Erfolg des Systems deutlich erhöhen. Auf den ersten Blick wirkt dies aufwändig und vielleicht sogar abschreckend. Dennoch gilt hier das Motto „Alles ist besser als nichts!“. Bereits einige wenige informelle Gespräche mit zukünftigen Benutzer:innen, Notizen über deren Arbeitsweise und Fotos von deren Arbeitsumgebung können schon helfen, um den Nutzungskontext erheblich besser zu verstehen. Selbst kleine Schritte sind immer noch besser, als dass eine Software vom bequemen Schreibtisch und fernab vom realen Nutzungskontext komplett am Bedarf vorbei gestaltet wird. Deshalb ist es auf jeden Fall wichtig, direkt mit den realen Zielgruppen in Kontakt zu treten - selbst wenn es nur für ein informelles Gespräch beim Kaffee ist.

Die vier Bestandteile des Nutzungskontextes im Detail

Die folgenden Aspekte und Methoden können als Inspiration für ein Vorgehen bei Ihrem System dienen. Sie können je nach System, Zielen und verfügbaren Ressourcen unterschiedlich relevant und umsetzbar sein.

1. Benutzer:innen

Die Benutzer:innen, also die reale Zielgruppe des Systems, stehen im Mittelpunkt der Gestaltung. Dieses Vorgehen nennt man auch benutzerzentrierte Gestaltung. Weiteres zu diesem Vorgehen erfahren Sie detailliert im Kapitel [Mensch-Computer-Interaktion](#)

Dabei können verschiedene Aspekte analysiert werden:

- **Fähigkeiten**, z.B. technisches Verständnis, Sprachkompetenz
- **Vorerfahrung und Wissen**, z.B. Umgang mit KI-Tools wie Chatbots oder Bildgeneratoren
- **Erwartungen**, z.B. durch vorherige Nutzung bestimmter Systeme
- **Demografische Daten**, wie Alter, Geschlecht, körperliche Einschränkungen
- **Mentale Eigenschaften**, z.B. Motivation, Einstellung gegenüber KI, Lernstil

2. Aufgaben

Die Aufgaben beschreiben, was die Benutzer:innen mit einem System erreichen wollen. Das hat großen Einfluss darauf, welche Funktionen ein System besitzt.

Dazu könnte man sich folgende Aspekte ansehen:

- **Aufgabenziel**, z.B. Informationen finden, Texte generieren
- **Einbettung in Arbeitsabläufe**, z.B. Teil eines Prozesses
- **Vorgaben und erwartete Ergebnisse**, z.B. qualitativ hochwertiger Textentwurf
- **Häufigkeit und Bearbeitungsdauer** von Aufgaben
- **Physische und mentale Anforderungen**, z.B. Konzentrationsaufwand

Im Rahmen der Entwicklung eines neuen Systems lohnt es sich, alle diese Aspekte der Aufgaben zu analysieren und zu dokumentieren.

Damit lassen sich klarere Anforderungen an das neue System formulieren.

Zum Beispiel kann die Häufigkeit und Bearbeitungsdauer einer Aufgabe eine große Rolle spielen. Wenn eine Aufgabe sehr typisch und häufig ist, so sollte die entsprechende Funktion zu deren Bearbeitung mit dem System besonders sorgfältig gestaltet sein!

Wird sie z.B. täglich oder sogar mehrmals täglich verwendet, dann ist unbedingt zu vermeiden, dass viele Informationen immer neu eingegeben oder Teilschritte stupide wiederholt werden müssen. Um häufige Aufgaben schneller bearbeiten zu können, dürfen auch Funktionen integriert werden, die etwas komplexer zu erlernen und zu bedienen sind, z.B. Tastaturkürzel, vorausgefüllte Vorlagen und Template oder die Möglichkeit zur Speicherung wiederverwendbarer Standard-

Textbausteine oder Makros. Das ist wünschenswert, sofern sich damit im Endeffekt die Aufgaben schneller und besser erledigen lassen. Man spricht hier vom „**Ease of Use**“ und die Effizienz steht dabei im Vordergrund, da man davon ausgeht, dass die Benutzer:innen in solchen Funktionen geschult werden und viel Übung haben.

Eine Funktion, die von ungeschulten Benutzer:innen nur sehr selten verwendet wird, muss sich dagegen viel einfacher, selbsterklärender und „intuitiver“ präsentieren. Man spricht hier vom „**Ease of Learn**“ und die Erlernbarkeit steht dabei im Vordergrund.

Ease of Use (auch Benutzungseffizienz) und **Ease of Learn** (oft als Erlernbarkeit bezeichnet) sind zwei zentrale, aber unterschiedliche Qualitätsmerkmale bei der Gestaltung und Bewertung von Softwaresystemen. Die Unterscheidung ist gerade im Hinblick auf unterschiedliche Zielgruppen, etwa interne Nutzer:innen (Mitarbeitende) vs. externe Nutzer:innen (Kunden:innen, Klienten), entscheidend.

Kriterium	Ease of Learn (Erlernbarkeit)	Ease of Use (Benutzungseffizienz)
Zielgruppe	Erstnutzende, Gelegenheitsnutzende	Routine-Nutzende, Experten:innen
Wichtig bei	Öffentlichen, selten genutzten Systemen	Internen, häufig genutzten Systemen
Fokus	Schnelle Einarbeitung, geringe Hürden	Effizienz, Produktivität, Zufriedenheit
Typische Metrik	Zeit bis zur ersten erfolgreichen Nutzung	Geschwindigkeit, Fehlerfreiheit, Zufriedenheit

Beispiele wie Ease of Use und Ease of Learn Einfluss haben

In einer Organisation zur Hilfe von Geflüchteten wird ein KI-gestütztes Übersetzungstool eingesetzt. Die Helfer:innen haben generell sehr unterschiedliche technische Fähigkeiten und Vorerfahrungen mit KI-Tools. Das System sollte daher neue oder ungeübte Nutzer:innen nicht mit zu vielen Funktionen überfordern. Funktionen für Fortgeschrittene, die eher erfahrenen Benutzer:innen helfen, sollten optional sein und den Erstkontakt nicht unnötig erschweren. Für die Geflüchteten muss das Tool zusätzlich auch verschiedene Alphabetisierungsniveaus abdecken. Manche Benutzer:innen werden nicht oder nur schlecht lesen können, weshalb hier eine Sprachausgabe und Piktogramme notwendig sind. Das Übersetzungstool muss also für viele unterschiedliche Bedürfnisse angepasst werden, um wirklich gebrauchstauglich zu sein.

Bei internen Systeme, z. B. für Mitarbeitende, kann es sinnvoll sein, den Ease of Learn etwas zugunsten des Ease of Use zu vernachlässigen, wenn die Nutzer:innen regelmäßig und intensiv mit dem System arbeiten. Eine kurze, initiale Schulung ist oft akzeptabel, solange das System anschließend effizient und produktiv genutzt werden kann.

Bei Front-facing Systeme (für Klienten/Kunden) hingegen ist der Ease of Learn besonders kritisch, weil die Nutzer:innen oft keine oder nur wenig Vorerfahrung mit dem System haben und keine Schulung erhalten. Das System muss somit schnell verständlich sein, damit sie ihre Ziele ohne Frustration erreichen.

3. Chatbots als Allheilmittel

Durch den Erfolg von ChatGPT wird heute sehr oft die Automatisierung von Aufgaben durch Chatbots als eine Art Allheilmittel betrachtet. Dies erscheint zunächst vielversprechend, da Chatbots bereits effizient und vielfältig eingesetzt werden, z.B. für die automatische Beantwortung von Kundenanfragen oder für Auskünfte über interne Prozesse in Organisationen oder gesetzliche Richtlinien. Hier ist jedoch Vorsicht geboten! Erfolgreiche Chatbots benötigen eine Datengrundlage mit entsprechend hoher Qualität. Die dort enthaltenen Informationen müssen korrekt und auch in für die KI verarbeitbaren Strukturen vorliegen. Chaotische, unstrukturierte und schlecht zu verarbeitende Sammlungen von Dateien und Dokumenten in verschiedenen Ordnerstrukturen und Formaten lassen sich auch durch eine moderne KI nicht einfach in eine zuverlässige Informationsquelle verwandeln. Auch hier gilt der Informatik-Grundsatz „**Garbage in, Garbage out!**“.

Selbst bei gepflegten Dokumentenbeständen als Wissensgrundlage für Chatbots zeigt sich nach der Implementierung nicht selten eine große Unzufriedenheit der Nutzer:innen, da ihre realen Aufgaben, Erwartungen, Informationsbedürfnisse und Anforderungen bei der Entwicklung nicht ausreichend berücksichtigt wurden. Das ist insbesondere dann problematisch, wenn die Qualität des eigenen Chatbots im Vergleich zu den mit dem gesamten Wissen der Welt trainierten KI-Systemen globaler IT-Konzerne in der Wahrnehmung der Benutzer:innen schlecht abschneidet.

Um solche Szenarien zu vermeiden, ist es ratsam, vorher den Nutzungskontext genau zu analysieren und darauf zu achten, dass die Aufgaben und Anforderungen der Benutzer:innen überhaupt mit den vorhandenen Technologien und Daten erfüllt werden können. Im Modul **Automatisierungspotenziale erkennen** werden wir uns eingehend damit

beschäftigen, wie Automatisierungs- und KI-Potenziale identifiziert werden können und wie man frühzeitig bewerten kann, welche Prozesse sich dafür eignen, bevor eine Umsetzung durchgeführt oder beauftragt wird.

4. Umgebung

Ein System muss je nach Umgebung sehr unterschiedliche Anforderungen erfüllen, um wirklich wirksam nutzbar zu sein. Die folgenden Bereiche gehören zur Umgebung

Soziale Umgebung:

- **Arbeitsstruktur**, z.B. Einzelarbeit oder Teamarbeit mit KI-Tools
- **Unterbrechungen** im Arbeitsalltag und bei Aufgaben
- **Unterstützung oder Schulungen**, z.B. IT-Support bei Problemen mit der KI oder Training mit neuen Systemen
- **Organisationskultur**, z.B. Akzeptanz von KI-Assistenz

Physische Umgebung:

- **Arbeitsplatzbedingungen**, z.B. Lärmpegel in der Umgebung und die Anwesenheit anderer Personen vs. Zuverlässigkeit und Datenschutz bei Spracheingabe, Spiegelungen, Lichteinstrahlung und Handschuhe vs. Outdoor-Einsatz eines Touchscreens
- **Technische Ausstattung**, z.B. verfügbare Rechen- und Grafikleistung auf Mobilgeräten, Stabilität und Geschwindigkeit der Internetverbindung, Bildschirmgröße und -auflösung

- **Arbeitsplatzausstattung**, z.B. Verwendung im Büro, mobile Verwendung im Stehen oder Sitzen, Verfügbarkeit von Teamarbeitsplätzen mit Projektor oder digitalem Whiteboard

Hier ist ein Beispiel: Eine KI-Anwendung zur Sprachübersetzung wird in zwei gemeinwohlorientierten Organisationen eingesetzt.

1. Mobiler Beratungsbus für Geflüchtete

In einem mobilen Beratungsbus für Geflüchtete arbeiten die Mitarbeitenden mit Tablets. Die Beratung findet häufig unter freiem Himmel statt, bei instabiler Internetverbindung und unter hohem Zeitdruck. Die Arbeit erfolgt dabei oft im Stehen oder in improvisierten Sitzpositionen.

In diesem Nutzungskontext ist der Einsatz heute typischer KI-Technologien, z. B. Chatbots, nur bedingt geeignet, da die technische Infrastruktur sowie die körperlichen und zeitlichen Rahmenbedingungen für deren Einsatz nicht optimal sind. Heutige Chatbots brauchen in der Regel schnelle Internetverbindungen und basieren auf Eingaben über Tastatur oder Sprache, die mit spürbaren Verzögerungen durch weit entfernte Server erst verarbeitet und dann beantwortet werden. Die Anwendung sollte dagegen Offline-Funktionalität und schnelle Spracheingabe berücksichtigen.

2. Digital ausgerichtete NGO mit Online-Beratung

In einer digital ausgerichteten NGO, die Online-Beratung anbietet, arbeiten die Teams in ruhigen Büros mit ergonomischen Arbeitsplätzen. Dort stehen leistungsfähige Laptops oder Desktop-PCs mit großen Bildschirmen, Headsets und stabiler Internetverbindung zur Verfügung. Die Beratungen folgen einem festen Zeitplan und ermöglichen konzentriertes Arbeiten.

5. Hilfsmittel

Die technischen Bedingungen umfassen folgende Punkte:

- **Hardware**, z.B. Headsets für Sprachassistenten, Tablets für mobile KI-Apps
- **Software**, z.B. spezialisierte KI-Tools, Betriebssysteme
- **Dokumentationen**, z.B. Handbücher, Tutorials
- **Nicht-digitale Werkzeuge und Objekte**, z.B. Stifte, Notizzettel, Formulare, Laufkarten

Obwohl Hardware, Software und Dokumentation natürlich einen unmittelbaren Einfluss auf die erfolgreiche Verwendung bzw. Usability eines Systems haben, sollte die Rolle der nicht-digitalen Werkzeuge und Objekte keinesfalls unterschätzt werden. Viele Prozesse laufen heute noch papier-basiert ab. Beispielsweise werden auch heute noch wichtige Informationen auf Handzetteln ausgegeben oder über Formulare abgefragt. Laufkarten und Notizen begleiten Personen, Gegenstände oder Akten durch eine Organisation und ermöglichen es, entscheidende

(Meta-)Informationen festzuhalten (z.B. Hintergründe, Internas, Vorgeschichten), die in digitalen Systemen oftmals komplett wegfallen. Es lohnt sich also einen Blick auf diese wichtigen nicht-digitalen Werkzeuge und Objekte zu werfen, bevor man Aufgaben digitalisiert oder automatisiert. Nur so kann sicher gestellt werden, dass den Benutzer:innen nachher nicht im System entscheidende Möglichkeiten zur Informationseingabe und -weitergabe fehlen.

Wie erfolgt eine Nutzungskontextanalyse?

Die Nutzungskontextanalyse untersucht die vier oben genannten Aspekte. Dabei ist zu unterscheiden, ob ein neues Produkt entwickelt oder ein bestehendes überarbeitet wird. Bei bestehenden Produkten kann auf vorhandenes Wissen oder frühere Tests zurückgegriffen werden. Bei Neuentwicklungen muss der Nutzungskontext vollständig neu erhoben werden.

Anmerkung: Dabei wird vom Idealfall ausgegangen. Wir wissen sehr gut, dass die Realität oft weit davon entfernt ist und eine solche Analyse oftmals nicht so umfassend durchgeführt werden kann. Daher gilt hier wieder „Alles ist besser als nichts!“.

Die Analyse erfolgt in zwei Schritten:

1. Datenerhebung

Zunächst werden Informationen über Benutzende, Aufgaben und Nutzungssituation gesammelt. Mögliche Methoden:

- **Interviews:** Gezielte Befragung von Nutzer:innen oder Expert:innen.
- **Umfragen:** Standardisierte Erhebung von Meinungen und Erfahrungen über Fragebögen.
- **Fokusgruppen:** Gruppendiskussion zur Sammlung unterschiedlicher Perspektiven.

2. Analyse & Dokumentation

Die gesammelten Daten werden geordnet und aufbereitet, beispielsweise durch:

- **Personas:** Fiktive, aber realitätsnahe Nutzerprofile.
- **Nutzungsszenarien, Problemszenarien:** Beschreibungen typischer Nutzungs- oder Problemsituationen.
- **Aufgabenanalyse:** Zerlegung von Aufgaben in Teilaufgaben (z.B. mittels Hierarchical Task Analysis).

Auf Basis dieser Analysen werden **Usability-Ziele** und **konkrete Nutzungsanforderungen** abgeleitet. Diese sind die Grundlage für Gestaltung und Evaluation. Später können damit gezielte **Usability-Tests** durchgeführt werden. Kontexte können sich ändern, also kann die Analyse kontinuierlich wiederholt werden. Mehr dazu in den Kapiteln **Mensch-Computer-Interaktion** und **Evaluationen**.

Ausblick

Die festgelegten Anforderungen und Ziele an das System beziehen sich nicht nur auf die Gebrauchstauglichkeit (Usability), sondern auch auf **emotionale Aspekte**. Dazu gehört, dass sich die Nutzung eines Systems stimmig, angenehm oder motivierend anfühlt. Mehr dazu erläutern wir im nächsten Kapitel zur **User Experience (UX)**.

04 User Experience

[Kursübersicht](#) > [UX und Usability](#)

User Experience umfasst das gesamte Nutzungserlebnis eines Produkts, einschließlich subjektiver Empfindungen wie Ästhetik, Vertrauen und Freude. UX geht über Usability hinaus und ist heute ein Schlüssel zum Erfolg aller Arten von Produkten.

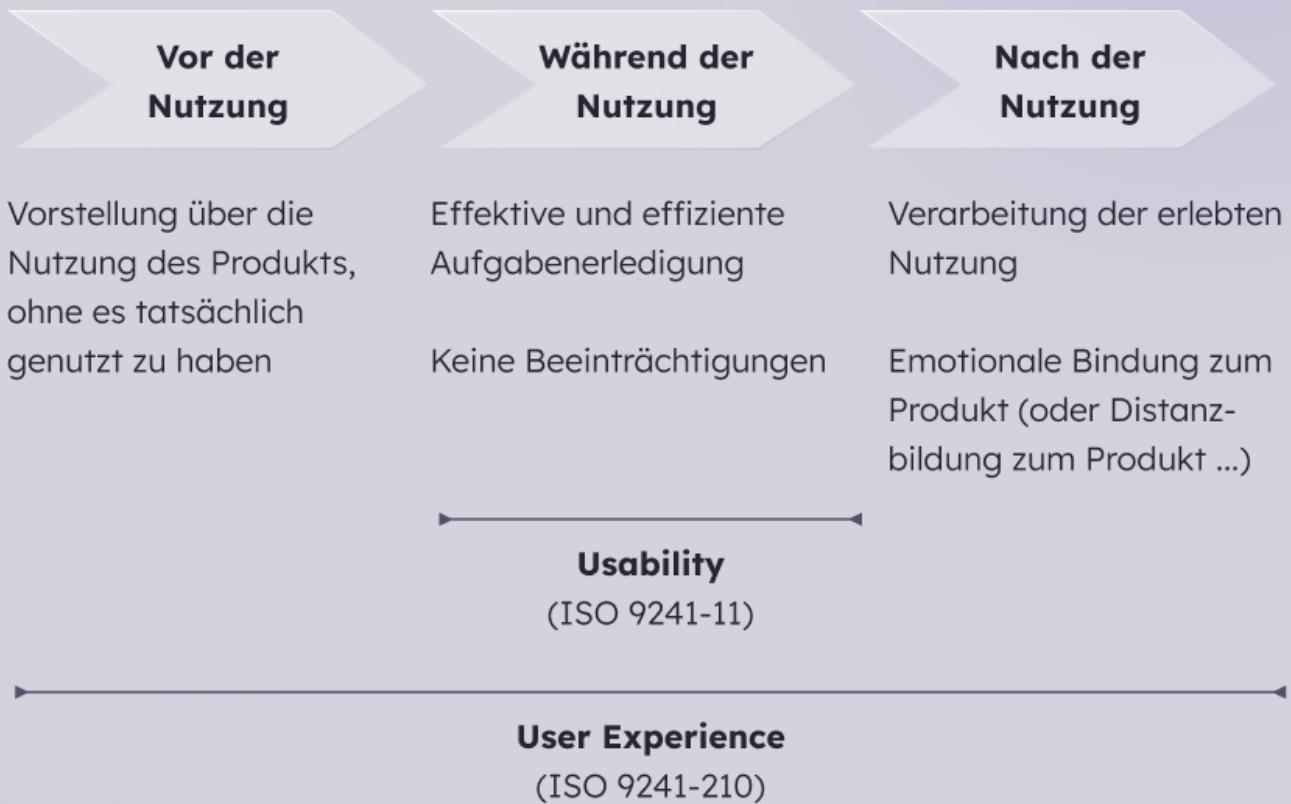
Was ist UX?

Die **User Experience** (UX) ist ein immer wichtigeres Ziel bei der Gestaltung digitaler Systeme. Auch wenn es dafür unterschiedliche Übersetzungen gibt, ist „**Benutzungserlebnis**“ unserer Meinung nach der passendste deutsche Begriff. UX beschreibt die Gesamtheit der Eindrücke, Gefühle und Erlebnisse, die Nutzer:innen **vor, während und nach der Interaktion** mit einem Produkt oder System haben. Gute Usability wird mittlerweile vorausgesetzt. Aber Erlebnisse, die den Benutzer:innen einen Eindruck von Leistungsfähigkeit, Eleganz, Attraktivität und Zuverlässigkeit des Systems vermitteln? Diese bleiben in Erinnerung und machen den Unterschied!

Nehmen wir als Beispiel ein Smartphone:

- Bereits vor der ersten Nutzung können Vorfreude und Erwartungen entstehen, etwa durch Werbung, Design oder das Markenimage.
- Während der Nutzung sind nicht nur Funktionalität und Usability wichtig, sondern auch der Spaß, die emotionale Wirkung und die Attraktivität des Systems bei der Verwendung.
- Nach der Nutzung, etwa nach der Verwendung einer App, dem Fotografieren oder einem Telefonat, können Gefühle wie Zufriedenheit, Vertrauen, Stolz, aber auch Frustration auftreten.

User Experience nach ISO



Die UX wird durch emotionale, soziale und kulturelle Aspekte geprägt, also dadurch, wie das Produkt in das Leben der Nutzer:innen

eingebettet ist. Ein Produkt, das nicht nur funktioniert, sondern auch Freude bereitet und positiv im Gedächtnis bleibt, hebt sich deutlich von anderen ab.

Laut ISO 9241-210 umfasst die UX „alle Wahrnehmungen und Reaktionen einer Person, die sich aus der Nutzung oder erwarteten Nutzung eines Produkts, Systems oder einer Dienstleistung ergeben.“

UX als Markenversprechen

Die Bedeutung von UX ist bei Systemen für gemeinwohlorientierte Organisationen nicht immer auf den ersten Blick ersichtlich. Hier steht meist nicht der Spaß im Vordergrund, sondern dass Endnutzer:innen gut beraten werden und relevante Informationen erhalten oder Mitarbeitende schnell zuverlässige Prognosen für ihre Ressourcenplanung bekommen.

Dennoch sind UX und ihre emotionalen Komponenten entscheidend. Jedes System, das im Kontext einer gemeinwohlorientierten Organisation angeboten wird, verändert deren Wahrnehmung von außen und auch von innen. Entspricht die Wirkung des Systems den selbstgewählten Mission, dem Selbstverständnis und der Zuverlässigkeit, für die die Organisation stehen will?

Wenn die Vorfreude oder Erwartungshaltung der Endnutzer:innen beispielsweise enttäuscht wird, weil ein Chatbot nicht so hilfreich ist wie erwartet, führt das zu Frustration und Ablehnung. Unter Umständen kann sogar ein Imageschaden für die Gesamtorganisation entstehen. Auch wenn es ein internes Prognosewerkzeug versäumt, seinen

Nutzer:innen die vorhandenen Beschränkungen und Unsicherheiten in der Vorhersage klarzumachen, kann das schnell zu einem Vertrauensverlust und einer kompletten Ablehnung des Systems führen. Das ist besonders tragisch, wenn das System eigentlich gute Dienste leisten könnte, aber vielleicht nur in wenigen Situationen überfordert war.

Es lohnt sich also, digitale Systeme nicht nur durch eine rein funktionale und pragmatische Brille zu betrachten. Anstelle von „Funktioniert es?“ und „Wie gut kommen die Nutzer:innen damit zurecht?“ tritt zusätzlich die Frage „Wie wirkt es auf die Nutzer:innen und ist dies die erwünschte Wirkung?“.

Wie erreicht man eine gute UX?

Eine gute User Experience (UX) kann nicht direkt „designt“ werden, denn sie entsteht im subjektiven Erleben der Nutzer:innen. Oder wie Preece et al. (2015) treffend formulieren:

„You cannot design a user experience, only design for a user experience.“

Das bedeutet: UX-Designer:innen können ein System unter Berücksichtigung des Nutzungskontextes so gestalten, dass positive Nutzungserlebnisse möglichst wahrscheinlich werden. Entscheidend ist dabei ein nutzerzentrierter Ansatz, bei dem der Mensch und seine Bedürfnisse im Mittelpunkt stehen.

Im nächsten Kapitel zur **Mensch-Computer-Interaktion (HCI)** lernen Sie zentrale Methoden zur nutzerzentrierten Gestaltung kennen. Zuvor befassen wir uns jedoch mit den Zielen und grundlegenden Aspekten guter UX.

Zentrale Ziele einer guten UX können sein:

- **Produktive und einfache Interaktion**, die Nutzer:innen effizient ans Ziel bringt
- **Erfüllende und zufriedenstellende Nutzung**, die den Bedürfnissen entspricht
- **Positive Emotionen**, etwa das Gefühl, etwas „Cooles“, „Einzigartiges“ oder „Elegantes“ zu nutzen
- **Stärkere Markenbindung**, weil das Nutzungserlebnis positiv in Erinnerung bleibt

Wichtige UX-Faktoren im Detail

1. Erlebnispotential

Gute UX ermöglicht nicht nur das Erreichen eines Ziels, sondern schafft Erlebnisse. Wichtige Fragen dabei sind:

- Wie fühle ich mich bei der Nutzung?
- Wer kann ich durch das Produkt sein?
- Welche Erlebnisse ermöglicht mir das Produkt?

2. Pragmatische und hedonische Qualität (nach Hassenzahl)

Wir können also festhalten, UX bedeutet, wie Menschen eine Website, App oder ein Produkt erleben - und das betrifft nicht nur, **wie gut etwas funktioniert**, sondern auch, **wie es sich dabei anfühlt**. Also: Es geht sowohl um die praktische Seite (z. B. „Finde ich schnell, was ich suche?“) als auch um die **emotionale** (z. B. „Macht es mir Spaß, das zu benutzen?“ oder „Fühlt es sich angenehm an?“). Im Forschungskontext lässt sich dies als so genannte pragmatische und hedonische Qualitäten aufschlüsseln:

Pragmatische Qualität (PQ)	Hedonische Qualität (HQ)
Klar, unterstützend, nützlich, beherrschbar	Besonders, beeindruckend, aufregend, cool
Fokus: Gebrauchstauglichkeit & Nutzen	Fokus: Vergnügen & Wohlbefinden

3. Weitere Aspekte guter UX

Aspekt	Beschreibung
Emotionale Bindung & Vertrauen	Positive Emotionen fördern die Bindung und Vertrauen in das Produkt.
Konsistenz	Das "Benutzungserlebnis" sollte über alle Geräte hinweg stimmig sein, egal ob Smartphone, Tablet oder Desktop.
Barrierefreiheit & Inklusion	UX muss alle Nutzer:innen einbeziehen, auch z. B. Menschen mit Seh- oder motorischen Einschränkungen.
Feedback & Fehlervermeidung	Nutzer:innen sollten klares Feedback erhalten und vor möglichen Fehlern geschützt werden.
Interaktives Design & ständige Verbesserung	UX ist ein fortlaufender Prozess. Nutzerfeedback ist essentiell für kontinuierliche Optimierung.

Beispiel KI-Chatbot

Hier finden Sie je ein Beispiel für eine gute und eine schlechte User Experience bei einem KI-Chatbot:

Beispiel für gute UX

Ein KI-gestützter Chatbot kann eine positive User Experience bieten, wenn er sich intuitiv bedienen lässt, natürliche Sprache verwendet und klare Anweisungen gibt. Besonders überzeugend ist das Benutzungserlebnis, wenn der Chatbot Probleme schnell und präzise löst, sich an die individuellen Bedürfnisse und Präferenzen der Nutzer:innen anpasst und dabei eine freundliche, empathische Atmosphäre schafft.

Beispiel für schlechte UX

Negative UX entsteht hingegen, wenn es zu Missverständnissen kommt, etwa weil Benutzereingaben falsch interpretiert oder unklare Antworten gegeben werden. Auch mangelnde Flexibilität, also die Unfähigkeit, auf unerwartete oder komplexe Anliegen einzugehen, beeinträchtigt das Nutzungserlebnis. Lange Reaktionszeiten, unnötige Interaktionen und wiederholte Eingabeaufforderungen führen schnell zu Frustration, da sie den Dialog ermüdend und ineffizient wirken lassen.

Fazit: UX bedeutet Nutzerzentrierung

User Experience bedeutet, den Menschen und seine Bedürfnisse konsequent in den Mittelpunkt zu stellen. Dabei geht es nicht nur um ansprechende Oberflächen, sondern um die Gestaltung sinnvoller, angenehmer und benutzerfreundlicher Produkte, die positive Erlebnisse und Emotionen fördern.

Im nächsten Kapitel lernen Sie, wie mithilfe nutzerzentrierter Methoden im Rahmen der Mensch-Computer-Interaktion (HCI) eine positive UX systematisch gestaltet und evaluiert werden kann.

Mensch-Computer-Interaktion

Kursübersicht > [UX und Usability](#)

Mensch-Computer-Interaktion ist eine interdisziplinäre Forschungsdisziplin, die sich unter anderem mit der Gestaltung von Interaktionen zwischen Menschen und Technologie beschäftigt. Ziel ist es, positive Nutzungserlebnisse durch optimierte Design- und Entwicklungsprozesse zu schaffen.

Definition Mensch-Computer-Interaktion (HCI)

Die Mensch-Computer-Interaktion (engl. Human-Computer Interaction, HCI) ist ein interdisziplinäres Forschungsfeld, das sich mit der Gestaltung, Umsetzung und Bewertung interaktiver Computersysteme für die Nutzung durch Menschen befasst. Im Zentrum steht die Frage, wie Computer und Technologien so gestaltet werden können, dass sie für Nutzer:innen möglichst einfach, effizient, angenehm und sinnvoll bedienbar sind.

Ursprünglich richtete sich die Nutzung von Computersystemen an Expert:innen. Heute sind interaktive Technologien allgegenwärtig und müssen für alle Menschen einfach und angenehm nutzbar sein.

HCI vs. UX vs. Interaktionsdesign: Der Begriff HCI wird überwiegend im wissenschaftlichen Kontext verwendet und bezieht sich meist auf die Mensch-Computer-Interaktion als Forschungsdisziplin. Das Konzept der User Experience (UX) ist eines der zentralen Ergebnisse dieser Forschung und hat sich schnell in der Praxis verbreitet. In dieser Praxis versuchen Interaktionsdesigner:innen die UX systematisch zu verbessern. Man kann also sagen: Die HCI erforscht die Beziehung zwischen Mensch und Computer. Sie liefert neue Konzepte, Methoden, Prozesse und Technologien für das Interaktionsdesign. Das Ziel ist dabei, die UX realer Systeme in der Praxis zu verbessern, z.B. durch die Schaffung positiverer oder "besserer" Nutzungserlebnisse.

Warum ist menschzentrierte Gestaltung wichtig?

Ein zentrales Ziel menschzentrierter Gestaltung ist es, die Lücke zwischen den Vorstellungen der Nutzer:innen und der tatsächlichen Funktionsweise eines Systems zu überbrücken. Denn Nutzer:innen und Designer:innen haben oft unterschiedliche **mentale Modelle**, also innere Vorstellungen davon, wie ein System funktioniert bzw. funktionieren soll.

Mentale Modelle beruhen auf Erfahrungen, Intuitionen und Analogien. Jakob Nielsen (2010) beschreibt sie so: „A mental model is what the user believes about the system at hand.“ Es basiert auf Glauben, nicht auf Fakten. Sie helfen dabei, Komplexität zu reduzieren, indem sie Konzepte und deren Beziehungen vereinfacht abbilden. Damit erfüllen

seine wichtige Rolle, um unseren Alltag voller Technologien zu bestreiten. Viele Menschen können bspw. ihre PKWs oder Mobiltelefone bedienen, weil sie ein mentales Modell von unterschiedlichen Funktionsweisen haben. Ein vollständiges Wissen darüber, wie genau ein Elektromotor funktioniert ist nicht notwendig, um ein E-Auto zu bedienen.

Designer:innen oder Entwickler:innen verfügen meist über ein tieferes technisches Verständnis, wodurch sich ihre Modelle oft von denen der Nutzer:innen unterscheiden. Umso wichtiger ist nutzerzentriertes Design: „Design with users, not just for users.“ Gute Gestaltung folgt den mentalen Modellen der Nutzer:innen.

Ein Beispiel für ein mentales Modell

Viele Nutzer:innen gehen davon aus, dass ein KI-Chatbot „wie ein Mensch“ denkt und versteht. Ihr mentales Modell basiert auf natürlicher Sprache, schnellen Antworten und auf scheinbarem Verständnis. In Wirklichkeit verarbeitet der Chatbot jedoch statistische Wahrscheinlichkeiten und greift auf große Sprachmodelle zurück. Ein echtes Verständnis im menschlichen Sinn ist somit nicht gegeben. Dennoch ermöglicht dieses vereinfachte Modell eine einfache Nutzung: Fragen stellen, Antworten erhalten, Probleme lösen.

Das Design solcher Systeme greift diese mentalen Modelle also auf, um die Bedienung zu erleichtern. Es kann aber auch zu falschen Erwartungen und Frustration führen, wenn der Chatbot an seine Grenzen stößt. Daher müssen klare Hinweise auf die tatsächlichen Fähigkeiten und Begrenzungen gegeben werden, um realistische Erwartungen zu fördern bspw. indem irreführende Beschreibungen wie “das System muss überlegen” oder “der Bot antwortet” vermieden werden, um kein menschliches Denken zu suggerieren.

Vorgehensweise für menschzentrierte Gestaltung

In der Gestaltung nutzerfreundlicher Systeme gibt es verschiedene bewährte Vorgehensweisen, die auf anerkannten Normen basieren und dabei helfen, die User Experience der Nutzenden zu verbessern. Zu den wichtigsten zählen das **User-Centered Design (UCD)** und der **Human-Centered Design Prozess gemäß ISO 9241-210**.

Beiden gemein ist die Idee, dass man ein System selten sofort perfekt gestalten kann. Ein iteratives Vorgehen, also wiederholtes Entwerfen, Testen und Verbessern mit Nutzer:innen ist entscheidend, um ein System/Tool mit guter UX zu schaffen.

Die vorgestellten Vorgehensweisen sind idealisierte Modelle, die in der Praxis häufig angepasst werden müssen. Zeit- und Ressourcenbegrenzungen sowie individuelle Anforderungen führen dazu, dass selten jeder Schritt vollständig umgesetzt werden kann. Dennoch bieten sie eine wertvolle Orientierung für eine nutzerzentrierte Gestaltung.

Um einen groben Überblick von den Vorgehensweisen zu bekommen, finden Sie im folgenden je eine kurze Zusammenfassung der wichtigsten Punkte, sowie des Ablaufs der einzelnen Vorgehensweisen. Für Ihr Projekt können Sie diese Vorgehensweisen als Inspiration oder Leitlinien nutzen, um einen für Sie passenden Prozess zu entwickeln.

1. User-Centered Design (UCD)

Der UCD ist ein iterativer Gestaltungsprozess, bei dem die Bedürfnisse, Erwartungen und Einschränkungen der Nutzer:innen von Anfang an in den Mittelpunkt gestellt werden.

Typische Schritte sind:

1

Kontextanalyse bedeutet, dass Nutzer:innen mit einem System ihre Aufgaben erfolgreich bewältigen können.

2

Design beschreibt, dass nutzerzentrierte Gestaltung von Lösungen, die sich an den Bedürfnissen, Zielen und Kontexten aus der Analyse orientiert.

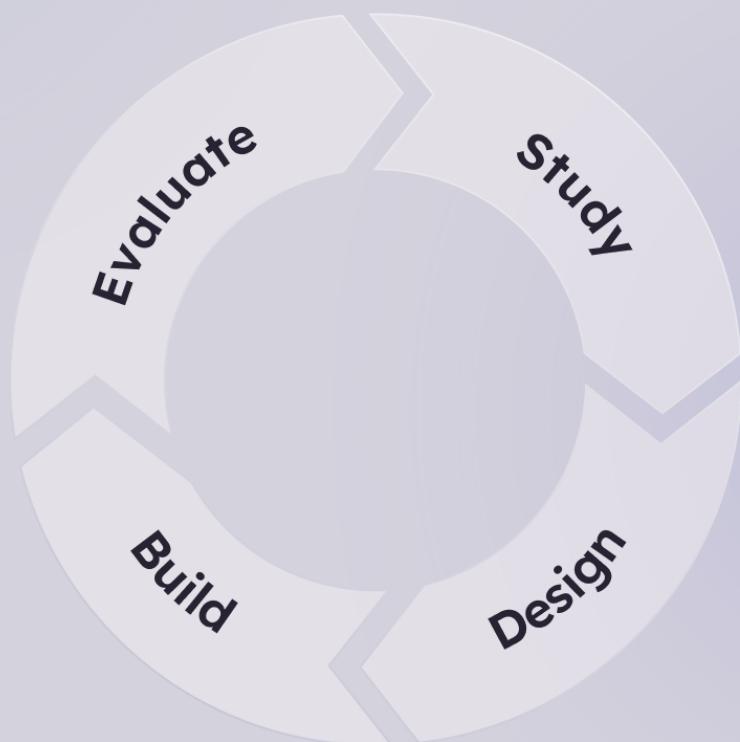
3

Implementierung (Build): Die entworfenen Lösungen werden technisch umgesetzt und in das reale System integriert.

4

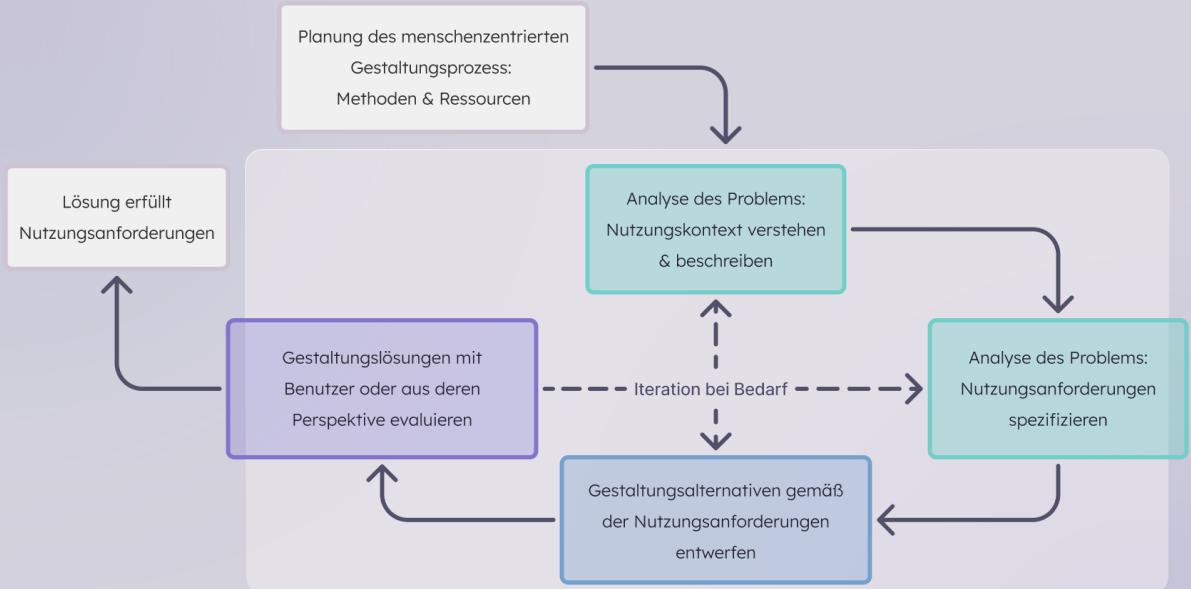
Evaluationen: In Usability-Tests, A/B-Tests oder durch reale Nutzungsszenarien („in the wild“) wird überprüft, wie gut das System tatsächlich funktioniert. Mehr dazu im Kapitel Evaluation.

Im letzten Schritt des iterativen Prozesses (understand) werden die Ergebnisse der Evaluationen interpretiert und die Erkenntnisse in die Study-Phase zurückgeführt.



2. Human-Centered Design (ISO 9241-210)

Der Human-Centered Design-Prozess (HCD) ist ein international genormtes, menschenzentriertes Gestaltungsmodell. Er legt besonderen Wert auf ein systematisches Vorgehen und definiert sechs zentrale Prinzipien, bei denen Nutzer:innen früh und kontinuierlich einbezogen werden für ein gesamtheitliches Systemverständnis.



Der Prozess beginnt mit dem Verstehen und Festlegen des Nutzungskontexts, leitet daraus Anforderungen an das System ab, entwickelt Lösungen iterativ und bewertet sie kontinuierlich mit Nutzer:innen, bis die Anforderungen erfüllt sind. Dabei sind alle Schritte eng verzahnt und wiederholen sich zyklisch zur stetigen Verbesserung der Gebrauchstauglichkeit.

Fazit

1

HCI zielt auf nutzerfreundliche Technik durch Gestaltung, die an den Bedürfnissen und mentalen Modellen der Nutzer:innen ausgerichtet ist.

2

Modelle wie UCD und HCD bieten strukturierte, iterative Prozesse für nutzerzentriertes Design.

3

Mentale Modelle vereinfachen Technik, müssen aber durch klares Design realistisch unterstützt werden.

06

Evaluation

[Kursübersicht](#) > [UX und Usability](#)

Als Evaluation bezeichnet man die Überprüfung und Messung von Usability und UX eines Systems. Durch Methoden wie Tests und Nutzerfeedback wird ermittelt, ob die gewünschte Nutzungserfahrung erreicht wird.

Warum sind Evaluationen wichtig?

Evaluationen sind, wie Sie im vorherigen Modul gesehen haben, ein wichtiger Bestandteil des benutzerzentrierten Gestaltungsprozesses. Sie helfen dabei sicherzustellen, dass ein System nicht nur funktional ist, sondern auch den Bedürfnissen der Nutzer:innen entspricht. Ziel ist es, zu überprüfen, ob die Nutzeranforderungen erfüllt werden, ob das System benutzerfreundlich ist und ob es gerne genutzt wird.

Dabei gilt: Usability-Tests können die Benutzerfreundlichkeit zwar verbessern, ersetzen aber keine gute Gestaltung. Eine gute User Experience lässt sich nicht einfach „herantesten“. Eine erfolgreiche

Evaluation setzt daher ein grundlegendes **Verständnis der Nutzer:innen** und **ihrer Aufgaben im Nutzungskontext** voraus. Nur so können Rückschlüsse gezogen und Verbesserungen erzielt werden.

Wann wird evaluiert?

Evaluationen sind nicht nur eine abschließende Qualitätskontrolle, sondern begleiten idealerweise den gesamten Entwicklungsprozess. Je nach Zeitpunkt unterscheiden sich die Ziele und Vorgehensweisen:

- 1.** Die **formative Evaluation** erfolgt während der Entwicklung. Ihr Ziel ist es, frühzeitig Schwächen und Usability-Probleme zu erkennen, um gezielt Verbesserungen vorzunehmen. Häufig kommt hier ein Prototyp zum Einsatz, der mit echten Nutzer:innen getestet wird.
- 2.** Die **summative Evaluation** findet hingegen am Ende des Entwicklungsprozesses statt. Hier geht es darum, die Gesamtqualität des Systems zu bewerten, zum Beispiel durch den Vergleich verschiedener Versionen, um herauszufinden, welche besser funktioniert.

Evaluationsart	Wann?	Ziel	Beispiel
Formative Evaluation	Während der Entwicklung	Probleme frühzeitig erkennen und verbessern	Usability-Test mit einem Prototyp
Summative Evaluation	Nach der Entwicklung	Gesamtqualität beurteilen, Designs vergleichen	Vergleich von Version A und Version B

Wie wird evaluiert und mit wem?

Beim Evaluieren unterscheidet man dabei häufig zwischen empirischen und analytischen Methoden. Beide Ansätze haben unterschiedliche Schwerpunkte und eignen sich je nach Projektphase, Zielsetzung und verfügbaren Ressourcen:

- 1. Empirische Methoden** ermöglichen tiefere Einblicke in das **tatsächliche Benutzungserlebnis**. Beispiele dafür sind ein Usability-Test oder Interviews mit Bürger:innen, die prüfen, wie verständlich und zugänglich eine KI-gestützte Plattform für Sozialleistungen ist.

2. Analytische Verfahren lassen sich hingegen besonders in **frühen Phasen** oder bei **begrenzten Ressourcen** effizienter einsetzen. UX-Expert:innen führen beispielsweise eine Art Inspektion bzw. Begutachtung des Systems auf der Basis ihrer Expertise durch, um frühzeitig Barrieren für ältere Menschen in einem KI-basierten Beratungs-Chatbot zu identifizieren.

	Empirische Evaluation	Analytische Evaluation
Wer evaluiert?	Echte Nutzer:innen	Expert:innen (z.B. UX-Designer:innen, Usability-Spezialist:innen)
Ziel	Beobachtung des Nutzerverhaltens bei konkreten Aufgaben	Bewertung anhand von Regeln und Erfahrung, frühzeitig Probleme identifizieren
Methoden	Usability-Tests, Interviews, Beobachtungen, Fragebögen	Heuristische Evaluation (z.B. nach Nielsen), Cognitive Walkthrough
Vorteil	Realitätsnahe Ergebnisse	Schnell, günstig, keine Nutzer:innen nötig
Nachteile	Zeit- und Kostenintensiv	Weniger objektiv, oft weniger praxisnah

Was sind Evaluationskriterien?

Um die Qualität der User Experience oder Usability systematisch beurteilen zu können, braucht es klare Evaluationskriterien. Sie helfen dabei, Ergebnisse nachvollziehbar und objektiv zu bewerten. Je nach Projekt und Zielgruppe können unterschiedliche Schwerpunkte gesetzt werden. Typische Kriterien sind zum Beispiel:



Effektivität: Können Nutzer:innen ihre Aufgaben erfolgreich erledigen?

Beispiel: 90% der Aufgaben werden korrekt abgeschlossen.



Effizienz: Wie schnell und mühelos gelingt die Nutzung?

Beispiel: Aufgabe lässt sich in unter 30 Sekunden oder mit maximal 10 Klicks lösen.



Zufriedenheit: Wie empfinden Nutzer:innen die Nutzung?

Beispiel: Keine negativen Rückmeldungen, positive Bewertungen.



Freude an der Nutzung: Macht die Anwendung Spaß oder fühlt sie sich frustrierend an?

Beispiel: Nutzer:innen berichten von einem positiven Erlebnis während der Nutzung.

Methoden zur Messung der Usability

Usability-Tests

In einem Usability-Test bearbeiten Nutzer:innen typische Aufgaben mit dem System, während Expert:innen ihr Verhalten beobachten, um Probleme und Barrieren zu erkennen. Diese Methode liefert Einblicke in die tatsächliche Nutzung und hilft dabei, konkrete Usability-Probleme zu identifizieren. Der Aufwand ist jedoch hoch, und der künstliche Testrahmen kann das Verhalten der Nutzer:innen beeinflussen. Zudem sind die Ergebnisse wegen der kleinen Teilnehmerzahl oft nicht vollständig repräsentativ.

Heuristische Evaluation

Bei der heuristischen Evaluation prüfen Expert:innen das System anhand von Usability-Heuristiken (z.B. die 10 Regeln von Nielsen). So lassen sich offensichtliche Usability-Probleme schnell identifizieren, ohne dass Nutzer:innen einbezogen werden müssen. Das macht die Methode kostengünstig und effizient für eine erste Einschätzung. Allerdings ist die Bewertung subjektiv und hängt stark von der Erfahrung der Expert:innen ab. Zudem werden oft nicht alle Probleme erkannt, die bei echten Nutzer:innen auftreten würden. Mehr zur Durchführung:

<https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/>

Fragebögen

Standardisierte Fragebögen sind eine einfache und effiziente Methode, um die subjektive Einschätzung der Usability durch Nutzer:innen zu erfassen. Zu bekannten zählen:

- die System Usability Scale (Deutsche Variante)
- Questionnaire for User Interface Satisfaction (QUIS)
- USE-Fragebogen (Usefulness, Satisfaction, and Ease of Use)

Diese Fragebögen lassen sich schnell durchführen, liefern quantitative Daten und können mit vielen Nutzer:innen gleichzeitig eingesetzt werden. Sie eignen sich besonders gut, um erste Usability-Einschätzungen zu gewinnen oder um mehrere Versionen eines Systems zu vergleichen.

Allerdings erfassen sie meist nur oberflächliche Aspekte und sind stark von der Ehrlichkeit und dem Verständnis der Teilnehmenden für die gestellten Fragen abhängig. Daher sollten sie idealerweise mit weiteren Methoden kombiniert werden, um ein umfassenderes Bild der Usability zu erhalten.

Methoden zur Messung der User Experience (UX)

Die User Experience umfasst mehr als reine Usability. Sie berücksichtigt auch Emotionen, Wahrnehmungen und das gesamte Nutzungserlebnis. Um diese Dimensionen zu erfassen, kommen ergänzende Methoden zum Einsatz:

Interviews

Interviews ermöglichen tiefe Einblicke in die Gedanken, Gefühle und Erfahrungen der Nutzer:innen. Durch offene Fragen erhält man detailliertes Feedback zur Wahrnehmung eines Produkts. Die Methode ist flexibel und kontextbezogen, jedoch zeitaufwendig und nur bedingt repräsentativ.

Emotionsanalyse und physiologische Messungen

Techniken wie Eye-Tracking, Emotionserkennung oder Hautleitfähigkeitsmessung erfassen unbewusste, emotionale Reaktionen auf ein Produkt. Diese objektiven Daten helfen dabei, subtile Aspekte der User Experience zu verstehen, die über bloße Befragungen hinausgehen. Sie sind aber oft technisch aufwändig, teuer, schwer zu interpretieren und können das Nutzerverhalten beeinflussen.

Fragebögen

Standardisierte Fragebögen wie der User Experience Questionnaire (UEQ) oder AttrakDiff liefern schnelle, vergleichbare Ergebnisse zu verschiedenen UX-Dimensionen. Sie sind einfach durchzuführen und gut validiert, erfassen jedoch hauptsächlich subjektive Eindrücke und bieten wenig Tiefe hinsichtlich der Ursachen hinter den Bewertungen.

Kombination von Methoden für eine umfassende Messung

Ein Multi-Methoden-Ansatz kombiniert verschiedene Erhebungsverfahren, um sowohl quantitative als auch qualitative Daten zu erfassen. So können unterschiedliche Perspektiven berücksichtigt und Stärken wie Schwächen eines Systems besser sichtbar gemacht werden.

Beispielsweise können Sie:

- 1. Zunächst einen Usability-Test** durchführen, um Beobachtungsdaten zu sammeln.
- 2. Anschließend einen Fragebogen** einsetzen, um subjektive Eindrücke zu messen.
- 3. Abschließend Interviews** führen, um Hintergründe und Details zu verstehen.

Allerdings wird auch hier von einem idealen Vorgehen ausgegangen und kann in realen Projekten den Anforderungen entsprechend abgewandelt werden.

Fazit

1

Evaluationen sind **kein Ersatz für gutes Design**, sondern ein Werkzeug zur Verbesserung.

2

Es gibt **verschiedene Arten von Evaluationen**, abhängig vom Zeitpunkt, Aufwand und Ziel: **Formativ vs. Summativ** oder **Empirisch vs. Analytisch**

3

Gute Evaluation basiert auf klarem Verständnis der Aufgaben, Nutzer:innen und Kontexte.

07

Fazit

Kursübersicht > UX und Usability

Die 5 Grundbegriffe und ihre Relevanz

1

Warum ist Usability wichtig für gemeinwohlorientierte KI?

Usability entspricht ungefähr dem, was viele unter „Benutzerfreundlichkeit“ verstehen. Sie stellt sicher, dass möglichst alle Menschen, unabhängig von ihrem technischen Wissen, die KI-Systeme effektiv, effizient und ohne große Hürden nutzen können. Dies ist entscheidend, um den Zugang zu gemeinwohlorientierten Diensten für eine breite Bevölkerungsschicht zu ermöglichen.

2

Warum ist es wichtig, den Nutzungskontext gemeinwohlorientierter KI zu analysieren und zu verstehen?

Die Analyse des Nutzungskontexts hilft, die Bedürfnisse, Fähigkeiten und Herausforderungen der Zielgruppen zu verstehen. Gerade in gemeinwohlorientierten Projekten sind Kenntnisse über soziale Hintergründe, Sprachkenntnisse, Ausbildungsniveau, Barrierefreiheit und die technische Ausstattung der Nutzer:innen essenziell, um die KI inklusiv und gerecht zu gestalten.

3

Welche Rolle spielt User Experience (UX) für gemeinwohlorientierte KI?

Eine positive UX fördert das Vertrauen in die KI und sorgt dafür, dass Menschen die Systeme gerne und regelmäßig nutzen. Gerade bei gemeinwohlorientierten Anwendungen kann eine intuitive und ansprechende UX die Akzeptanz und Wirksamkeit erheblich steigern.

4

Was hat Mensch-Computer-Interaktion mit gemeinwohlorientierter KI zu tun?

Die Mensch-Computer-Interaktion (MCI) oder Human-Computer Interaction (HCI) ist eine Forschungsdisziplin und liefert uns die notwendigen Methoden und Modelle, um die Interaktion zwischen Mensch und KI zu verbessern. Sie hilft, die Schnittstellen gebrauchstauglich, verständlich, aufgabenangemessen und barrierefrei zu gestalten. Besonders bei gemeinwohlorientierten Projekten ist es wichtig, die Bedürfnisse vielfältiger Nutzergruppen - etwa Senioren, Menschen mit Behinderungen oder technikferne Personen – zu berücksichtigen.

5

Wieso sollte man die UX gemeinwohlorientierter KI evaluieren?

Eine Evaluation macht sichtbar, ob die gemeinwohlorientierte KI tatsächlich ihren Zweck erfüllt und von den Menschen akzeptiert wird. Durch Tests, Nutzerfeedback und Datenauswertungen kann man sicherstellen, dass die Systeme inklusiv, effizient und nachhaltig sind. Evaluationen helfen außerdem dabei, mögliche Diskriminierungen durch die Bedienkonzepte oder die KI frühzeitig zu erkennen und zu beheben.

08

Quellen

Kursübersicht > [UX und Usability](#)

Literaturverzeichnis

Nutzungskontext

- Burmester, M. (2007). Usability und Design. In R. Schmitz (Hrsg.), *Kompendium Medieninformatik* (S. 245–302). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-36630-0_5
- DIN Media GmbH. (2018). *DIN EN ISO 9241-11:2018-11, Ergonomie der Mensch-System-Interaktion - Teil_11: Gebrauchstauglichkeit: Begriffe und Konzepte (ISO_9241-11:2018); Deutsche Fassung EN_ISO_9241-11:2018*. DIN Media GmbH.
<https://doi.org/10.31030/2757945>

User Experience

- Diefenbach, S., & Hassenzahl, M. (2017). *Psychologie in der nutzerzentrierten Produktgestaltung*. Springer Berlin Heidelberg.
<https://doi.org/10.1007/978-3-662-53026-9>

- Ecker, M. (2016). *Usability und Usability Engineering zur Gestaltung von Lernsystemen* (Technischer Bericht 1/2015, S. 38). Pädagogische Hochschule Weingarten AG Mediendidaktik und Visualisierung (MEVis). <https://hsbwgt.bsz-bw.de/frontdoor/index/index/year/2016/docId/191>
- Hartson, R., & Pyla, P. S. (2019). *The UX book: Agile UX design for a quality user experience* (Second edition). Morgan Kaufmann, an imprint of Elsevier.
- Hassenzahl, M. (2003). The Thing and I: Understanding the Relationship Between User and Product. In M. A. Blythe, K. Overbeeke, A. F. Monk, & P. C. Wright (Hrsg.), *Funology* (Bd. 3, S. 31–42). Springer Netherlands. https://doi.org/10.1007/1-4020-2967-5_4
- Hassenzahl, M., Burmester, M., & Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In G. Szwilus & J. Ziegler (Hrsg.), *Mensch & Computer 2003* (Bd. 57, S. 187–196). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-80058-9_19
- International Organization for Standardization. (2020). *DIN EN ISO 9241-210:2020-03, Ergonomie der Mensch-System-Interaktion—Teil 210: Menschzentrierte Gestaltung interaktiver Systeme (ISO 9241-210:2019); Deutsche Fassung EN ISO 9241-210:2019*. DIN Media GmbH. <https://doi.org/10.31030/3104744>
- Jetter, H.-C. (2006). Die MCI im Wandel: User Experience als die zentrale Herausforderung? In A. Münch, R. Oppermann, & M. Herczeg (Hrsg.), *Mensch und Computer 2006: Mensch und Computer im Strukturwandel* (S. 65–72). Oldenbourg Verlag.
- Jordan, P. W. (2000). *Designing Pleasurable Products* (0 Aufl.). CRC Press. <https://doi.org/10.4324/9780203305683>
- Norman, D. A. (2013). *The design of everyday things* (Rev. and expanded edition). MIT Press.

- Sharp, H., Rogers, Y., & Preece, J. (2019). *Interaction design: Beyond human-computer interaction* (Fifth edition). Wiley.

Mensch-Computer-Interaktion

- Constantine, L. L., & Lockwood, L. A. D. (1999). *Software for use: A practical guide to the models and methods of usage-centered design; [web powered]*. Addison Wesley.

- DIN EN ISO 9241-210:2020-03, *Ergonomie der Mensch-System-Interaktion - Teil 210: Menschzentrierte Gestaltung interaktiver Systeme (ISO 9241-210:2019); Deutsche Fassung EN ISO 9241-210:2019*. (2020). DIN Media GmbH.

<https://doi.org/10.31030/3104744>

- Ecker, M. (2016). *Usability und Usability Engineering zur Gestaltung von Lernsystemen* (Technischer Bericht 1/2015, S. 38). Pädagogische Hochschule Weingarten AG Mediendidaktik und Visualisierung (MEVis). **<https://hsbwgt.bsz-bw.de/frontdoor/index/index/year/2016/docId/191>**

- Harper, R., Rodden, T., Rogers, Y., & Sellen, A. (2008). *Being human: Human-computer interaction in the year 2020*. Microsoft Research. **<https://www.microsoft.com/en-us/research/project/being-human/>**

- Hartson, R., & Pyla, P. S. (2019). *The UX book: Agile UX design for a quality user experience* (Second edition). Morgan Kaufmann, an imprint of Elsevier.

- Nielsen, J. (2010). *Mental Models*.

<https://www.nngroup.com/articles/mental-models/>

- Preece, J., Rogers, Y., & Sharp, H. (2019). *Interaction design: Beyond human-computer interaction* (5. Aufl.). Wiley.

Evaluation

- Bogner, A., Littig, B., & Menz, W. (2014). *Interviews mit Experten: Eine praxisorientierte Einführung*. Springer Fachmedien Wiesbaden.
<https://doi.org/10.1007/978-3-531-19416-5>
- Hartson, R., & Pyla, P. S. (2019). *The UX book: Agile UX design for a quality user experience* (Second edition). Morgan Kaufmann, an imprint of Elsevier.
- Helfferich, C. (2009). *Die Qualität qualitativer Daten*. VS Verlag für Sozialwissenschaften. <https://doi.org/10.1007/978-3-531-91858-7>
- Kruse, J., & Schmieder, C. (2015). *Qualitative Interviewforschung: Ein integrativer Ansatz* (2., überarbeitete und ergänzte Auflage). Beltz Juventa.
- Nielsen, J. (1994, April 24). *10 Usability Heuristics for User Interface Design*. Nielsen Norman Group.
<https://www.nngroup.com/articles/ten-usability-heuristics/>

Modul:

02 KI-bezogene UX

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

1 Einleitung in KI-bezogene UX

Kursübersicht > KI-bezogene UX

In diesem Modul erläutert Tim Schrills konkretere UX-Aspekte, die beim Design und der Implementierung von KI-Systemen beachtet werden sollten.

KI-Systeme greifen tief in unsere Informationsverarbeitung ein, was zusätzliche Aspekte für die User Experience (UX) relevant macht. Diese neuen Dimensionen sind essentiell, da sie direkt beeinflussen, wie Menschen mit KI-Systemen interagieren und diese wahrnehmen. Mit der zunehmenden Integration von KI in alltägliche Systeme ist es entscheidend, KI-bezogene Aspekte in der UX zu berücksichtigen. Dazu werden in diesem Modul fünf wichtige Aspekte genannt und deren Relevanz für das Design und die Entwicklung von KI-Systemen erklärt.

Im folgenden Video wird ein Überblick über die unterschiedlichen Aspekte bei der Identifikation gegeben, auf die in den folgenden Kapiteln näher eingegangen wird.



<https://youtu.be/dbBIEqBMudI>

In der vorigen Lektion haben wir etablierte UX-Konstrukte betrachtet, die allgemeine Nutzererfahrungen beschreiben. KI-Systeme hingegen greifen tief in unsere Informationsverarbeitung ein, was zusätzliche Aspekte für die User Experience (UX) relevant macht. Diese neuen Dimensionen sind essentiell, da sie direkt beeinflussen, wie Menschen mit KI-Systemen interagieren und diese wahrnehmen.

Wichtige Aspekte der KI-bezogenen UX

1

Wahrgenommene Autonomie

Dieser Aspekt beschreibt, wie sehr Nutzende das Gefühl haben, selbstständig zu handeln und Entscheidungen zu treffen, während sie mit einem KI-System interagieren.

2

Wahrgenommenes Situationsbewusstsein

Dies bezieht sich auf das Verständnis der Nutzende über ihre Umgebung und die Änderungen, die durch das KI-System verursacht werden.

3

Wahrgenommene Mentale Belastung

Dieser Aspekt umfasst den kognitiven Aufwand, der erforderlich ist, um Informationen zu verarbeiten und Entscheidungen zu treffen, und die potenzielle Überlastung durch zu viele Informationen.

4

Wahrgenommene Vertrauenswürdigkeit

Damit ist das Vertrauen gemeint, das Nutzende in ein KI-System haben, basierend auf dessen Handlungen.

5

Wahrgenommene Diagnostizität

Dies beschreibt das Vertrauen der Nutzende in die Diagnosen oder Vorschläge des KI-Systems und wie gut diese die gewünschten Ergebnisse liefern.

Warum sind diese Aspekte wichtig für die Automation-Related UX?

Mit der zunehmenden Integration von KI in alltägliche Systeme ist es entscheidend, diese Aspekte in der UX zu berücksichtigen. KI-Systeme übernehmen Aufgaben, die früher Menschen vorbehalten waren, und erfordern daher, dass Nutzer ein Gefühl von Kontrolle, Verständnis und Vertrauen behalten. Die Gestaltung von KI-Systemen muss sicherstellen, dass Nutzer die Systeme nicht nur effektiv nutzen, sondern sich auch sicher und autonom fühlen können.

Beispiel: KI-gestütztes Verkehrsmanagementsystem

Stellen Sie sich ein KI-gestütztes Verkehrsmanagementsystem vor. Dieses System analysiert Verkehrsdaten und schlägt Optimierungen für Ampelphasen vor. Nutzer wie Verkehrsingenieure müssen das System verstehen, seine Vorschläge bewerten und gegebenenfalls anpassen können. Hier ist es wichtig, dass das System Transparenz bietet, um das Situationsbewusstsein zu fördern, eine angenehme Menge an Informationen bereitstellt, um Überlastung zu vermeiden, und Vertrauen in die Automatisierung aufbaut.

Was können Sie von dieser Lektion erwarten?

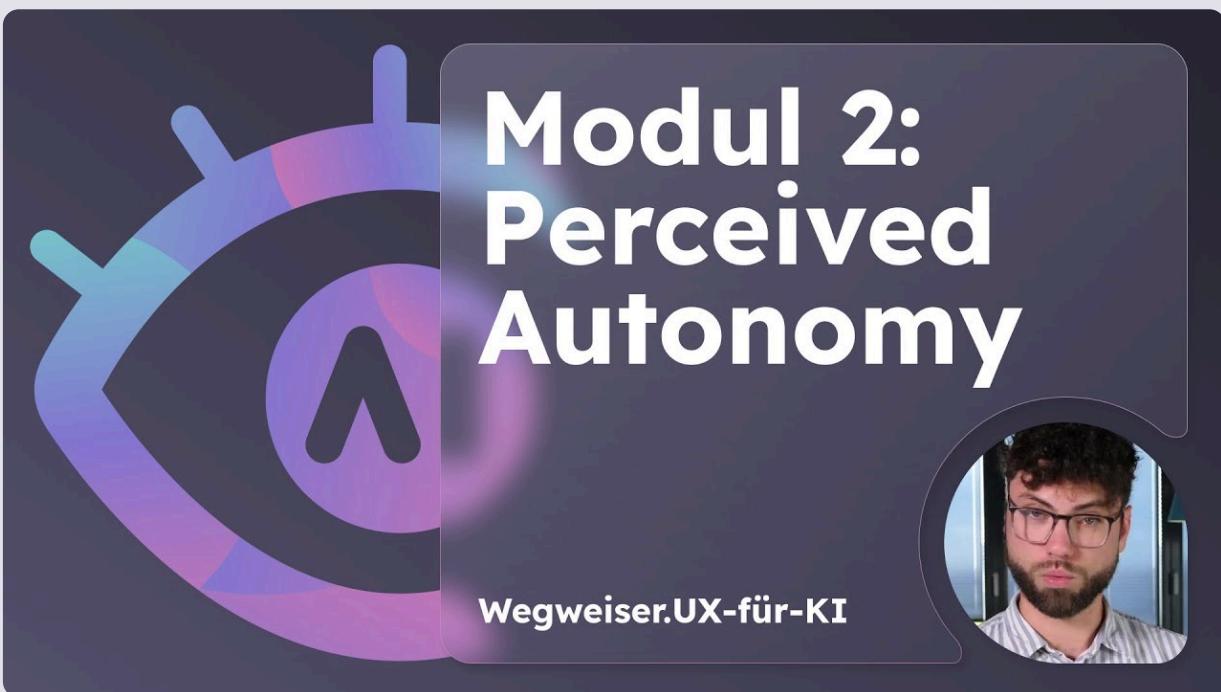
In den folgenden Kapiteln werden wir detailliert untersuchen, wie diese fünf Aspekte – Wahrgenommene Autonomie, Situationsbewusstsein, Mentale Arbeitsbelastung, Vertrauenswürdigkeit und Confidence/Diagnosticity – die Gestaltung und Nutzung von KI-Systemen beeinflussen. Wir werden sehen, wie diese Aspekte in der Praxis umgesetzt werden können und welche Designrichtlinien helfen, eine positive und effektive Mensch-KI-Interaktion zu fördern. Freuen Sie sich auf eine tiefere Auseinandersetzung mit den Herausforderungen und Chancen der KI-bezogenen UX!

Wahrgenommene Autonomie

Kursübersicht > KI-bezogene UX

Dieser Aspekt beschreibt, wie sehr Nutzende das Gefühl haben, selbstständig zu handeln und Entscheidungen zu treffen, während sie mit einem KI-System interagieren.

Im folgenden Video wird ein Überblick über den Begriff Wahrgenomme Autonomie gegeben.



<https://youtu.be/7QDwxEpQNHk>

1. Definition wahrgenommener Autonomie

Die **wahrgenommene Autonomie** beschreibt, inwiefern Nutzende eines KI-Systems das Gefühl haben, selbstständig handeln und Entscheidungen treffen zu können. Diese Wahrnehmung beeinflusst das Vertrauen in das System, die Interaktion damit und die Zufriedenheit der Nutzenden. Wesentlich ist, ob das System als transparent und unterstützend wahrgenommen wird oder ob es lediglich als ausführendes Werkzeug dient.

2. Relevante Konzepte und Modelle

Automatisierungsstufen (Levels of Automation)

Die **Automatisierungsstufen** (Levels of Automation, LOA) wurden entwickelt, um Abstufungen oder Kategorien der **Autonomie** zu veranschaulichen. Diese Struktur hilft zu verstehen, **wie Menschen mit automatisierten Systemen interagieren** und beschreibt, welche Aufgaben entweder vom Menschen oder von der Maschine übernommen werden.

Das LOA-Modell nach Parasuraman et al. (2000) umfasst **zehn Stufen der Aufgabenteilung** und Verantwortlichkeit zwischen Mensch und Maschine. Je nach Modell und Anwendungsbereich können die

Automatisierungsstufen jedoch variieren. So definiert beispielsweise die Society of Automotive Engineers (SAE) **fünf Stufen** der Automatisierung im Bereich **autonomer Fahrzeuge** (Hopkins & Schwanen, 2021).

Einfluss der Automatisierung auf psychologische Variablen

Die Einführung von Automatisierung hat einen signifikanten Einfluss auf psychologische Variablen wie **Arbeitsbelastung, Fähigkeiten, Vertrauen und Situationsbewusstsein** der Nutzer (Parasuraman et al., 2000).

Daher ist es entscheidend, das geeignete Automatisierungsniveau je nach Aufgabe auszuwählen, um unerwünschte Effekte zu vermeiden.

Vier Stufen der Informationsverarbeitung

Parasuraman, Sheridan und Wickens (2000) verknüpften die Automatisierungsstufen mit vier grundlegenden Funktionen, die auf einem Modell der menschlichen Informationsverarbeitung basieren und in einem Mensch-Maschine-System unterstützt werden sollen:

- 1.** Informationsbeschaffung,
- 2.** Informationsanalyse,
- 3.** Entscheidungsfindung
- 4.** Handlungsausführung.

Dieses Modell bietet eine strukturierte Herangehensweise zur Klassifizierung von Aufgaben, bei denen Automatisierung den Menschen unterstützen kann.

Fehler eines Systems in späteren Stadien können störender wirken als in Systemen, bei denen die Automatisierung höchstens bis zur Phase der Informationsanalyse eingesetzt wird. Akzeptanz, Vertrauen und Leistung können abnehmen, wenn in den späteren Phasen der Informationsverarbeitung zu viel Automatisierung vorhanden ist (Onnasch et al., 2014).

Einfluss der Automatisierung auf die menschliche Leistung

Parasuraman et al. (2000) haben vier zentrale Faktoren identifiziert, die beeinflussen, wie Automatisierung die menschliche Leistung beeinflussen kann:

- 1. Situationsbewusstsein:** Das Verständnis der aktuellen Umgebung und Situation durch den Menschen, welches durch Automatisierung entweder gefördert oder beeinträchtigt werden kann.
- 2. Vertrauen:** Das Vertrauen der Nutzenden in das System ist entscheidend für eine erfolgreiche Interaktion und hängt stark von der wahrgenommenen Autonomie ab.
- 3. Abbau von Fähigkeiten:** Hohe Automatisierung kann zu einem Rückgang menschlicher Fähigkeiten führen, da weniger manuelle Eingriffe und Entscheidungen nötig sind.
- 4. Arbeitsbelastung:** Automatisierung kann die Arbeitsbelastung entweder reduzieren oder erhöhen, abhängig davon, wie gut sie an die Bedürfnisse der Nutzenden angepasst ist.

Anpassbare vs. selbstanpassende Automatisierung

Die Wahl des Automatisierungs niveaus hat einen signifikanten Einfluss auf die Mensch-Automation-Interaktion. Es gibt zwei grundlegende Ansätze: **anpassbare Automatisierung** und **selbstanpassende Automatisierung**. In der anpassbaren Automatisierung wählt der Benutzer das Automatisierungs niveau manuell basierend auf seinen eigenen Bedürfnissen und Vorlieben. Im Gegensatz dazu überwacht die selbstanpassende Automatisierung den Zustand des Benutzers, wie Arbeitsbelastung oder Wachsamkeit, und passt das Automatisierungs niveau automatisch an.

1. Anpassbare Automatisierung

- Der Benutzer hat die Kontrolle über die Auswahl des Automatisierungs niveaus.
- Bietet Flexibilität und Anpassung an individuelle Präferenzen.
- Vorteilhaft, wenn Benutzer die Systeme nach ihren eigenen Bedürfnissen steuern wollen.

2. Selbstanpassende Automatisierung

- Das System überwacht den Benutzer (z. B. Arbeitsbelastung) und passt das Automatisierungs niveau automatisch an.
- Hilfreich in dynamischen Umgebungen, da es auf Veränderungen reagiert, ohne dass der Benutzer eingreifen muss.
- Kann die Arbeitsbelastung reduzieren, birgt jedoch das Risiko, dass der Benutzer die Kontrolle verliert.

3. Studien zur User Experience und KI

Kaber & Endsley (2004): The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task.

Untersuchten die Auswirkungen von adaptiver Automatisierung auf die Leistung des Menschen, das Situationsbewusstsein und die Arbeitsbelastung in dynamischen Umgebungen.

Niedriges Automatisierungsniveau: Verbessert die Leistung, da der Benutzer stark eingebunden bleibt (Kaber & Endsley, 2004). **Mittleres Automatisierungsniveau:** Führt zu verbessertem Situationsbewusstsein,

was entscheidend für komplexe und dynamische Aufgaben ist. Allerdings führen mittlere Automatisierungsstufen nicht immer zu besserer Leistung oder geringerer Arbeitsbelastung, obwohl sie das Situationsbewusstsein verbessern.

Rieger et al. (2022): Challenging presumed technological superiority when working with (artificial) colleagues.

Diese Studie untersucht, wie Menschen klassische Automatisierung und KI-basierte Systeme häufig als **Black Boxes** wahrnehmen, ohne signifikante Unterschiede in ihrer Wahrnehmung beider Technologien. Dies führt zu einem unvollständigen Automatisierungsschema, da Transparenz fehlt. Interessanterweise verändert sich die Präferenz der Menschen zwischen der Interaktion als Ratsuchender und als Bewerteter.

Während beim gemeinsamen Arbeiten menschliche Faktoren wie **Intuitivität** und **Fachwissen** bevorzugt werden, können bei der Bewertung durch Maschinen deren **Objektivität** und **Konsistenz** als vorteilhaft angesehen werden.

Deci & Ryan (1985): Self-determination theory (SDT)

Die Selbstbestimmungstheorie (Self-Determination Theory) definiert drei universelle grundlegende psychologische Bedürfnisse (Basic Psychological Needs, BPNs):

- 1. Autonomie:** das Gefühl, Kontrolle über eigene Entscheidungen und Handlungen zu haben
- 2. Kompetenz:** das Erleben von Wirksamkeit und Beherrschung einer Aufgaben
- 3. soziale Verbundenheit:** sich um andere zu kümmern und im Gegenzug Fürsorge zu erfahren

4. Operationalisierung: Fragebögen und Messinstrumente

Zoubir (2024): Preference for Automation Types Scale (PATS)

Ein Fragebogen zur Erfassung von Präferenzen der Nutzer hinsichtlich Automatisierungsaufgaben, basierend auf den Modellen von Parasuraman et al. (2000). Dieser misst, inwieweit Nutzer Automatisierung in verschiedenen Phasen der Informationsverarbeitung bevorzugen.

Moradbakhti et. al (2024): Basic Psychological Need Satisfaction for Technology Use (BPN-TU)

Die BPN-TU ist eine Skala zur Messung der Befriedigung grundlegender psychologischer Bedürfnisse bei der Nutzung von Technologie. Gemäß der Selbstbestimmungstheorie ist die Befriedigung der grundlegenden psychologischen Bedürfnisse nach Autonomie, Kompetenz und Verbundenheit entscheidend für das Wohlbefinden und die autonome Motivation.

5. Design-Guidelines für eine gute UX

1. Adaptive Automatisierung ermöglichen

Beispiel: Verkehrsmanagementsysteme für städtischen Verkehr

Ein städtisches Verkehrsmanagementsystem bietet verschiedene Automatisierungsstufen, wie z. B. die automatische Steuerung von Ampeln oder die manuelle Steuerung durch Verkehrsingenieure. Über eine benutzerfreundliche Oberfläche können die Verantwortlichen je nach Verkehrsaufkommen und speziellen Ereignissen die Automatisierungsstufe flexibel anpassen.

Diese Anpassungsmöglichkeit erlaubt eine präzise und flexible Steuerung des Verkehrsflusses, reduziert Staus und priorisiert den öffentlichen Verkehr. Die Nutzenden behalten dabei die Kontrolle und können den Automatisierungsgrad individuell anpassen, was zu einem effizienteren und reibungsloseren Verkehrserlebnis führt.

2. Situationsbewusstsein unterstützen

Beispiel: Notfallmanagementsysteme in Städten

Ein Notfallmanagementsystem liefert Echtzeitdaten zu städtischen Notfällen wie Verkehrsunfällen, Bränden oder Überschwemmungen. Es bietet automatisierte Empfehlungen für Evakuierungs Routen und Einsatzplanungen, die Einsatzleiter bei Bedarf manuell anpassen können.

Dank der Echtzeitinformationen können Einsatzkräfte schnell und präzise Entscheidungen treffen. Die Kombination aus automatisierten Vorschlägen und menschlichem Eingriff auf mittleren Automatisierungsstufen sorgt für ein optimales Gleichgewicht zwischen Effizienz und Sicherheit.

3. Flexibilität bei der Informationsverarbeitung

Beispiel: Umweltüberwachungssysteme in Städten

Ein Umweltüberwachungssystem erlaubt es den Nutzenden, die Art und Menge der überwachten Daten individuell festzulegen – etwa zur Überwachung der Luftqualität, Lärmelastung oder Wasserverschmutzung. Nutzende können den Fokus je nach Dringlichkeit und Prioritäten anpassen und die Detailtiefe der Analysen steuern.

Diese Flexibilität ermöglicht es den Nutzenden, auf spezifische Umweltfaktoren einzugehen und die Überwachung an aktuelle Bedürfnisse anzupassen. Dadurch wird eine maßgeschneiderte Umweltpolitik möglich, die effektiver auf akute Herausforderungen reagiert.

4. Transparenz sicherstellen

Beispiel: Medizinisches Diagnosetools

Ein KI-basiertes Diagnosetool für Ärzte zeigt nicht nur die Diagnoseergebnisse, sondern auch die zugrunde liegenden Daten und die Logik hinter der Entscheidung an. Die Entscheidungswege werden visualisiert, und das System erklärt, warum bestimmte Diagnosen vorgeschlagen wurden.

Diese Transparenz stärkt das Vertrauen der Ärzte in die KI, da sie genau nachvollziehen können, wie die Empfehlungen zustande kommen. Dies fördert einen effizienteren und informierten Entscheidungsprozess.

Beispiel: Überwachungssysteme für den öffentlichen Verkehr

Ein städtisches Verkehrssystem analysiert den Verkehrsfluss und erklärt transparent, wie Ampelschaltungen optimiert oder bestimmte Routen priorisiert werden.

Durch die klare Kommunikation der Algorithmen und Entscheidungsprozesse wird das Vertrauen der Öffentlichkeit gestärkt. Nutzende können nachvollziehen, wie Entscheidungen getroffen wurden, was ihre Akzeptanz und das Gefühl der Autonomie im Umgang mit dem System verbessert.

5. Nutzerzentrierte Anpassung

Beispiel: Smart City Mobilitätsplattformen

Eine Smart City Mobilitätsplattform ermöglicht es den Bürgern, ihre persönlichen Verkehrspräferenzen festzulegen - von bevorzugten Verkehrsmitteln über favorisierte Routen bis hin zu umweltbewussten Zielen wie der Reduzierung des CO2-Fußabdrucks. Die Plattform generiert daraufhin maßgeschneiderte Vorschläge, z. B. alternative Verkehrsmittel oder Fahrgemeinschaften, die den individuellen Präferenzen der Nutzenden entsprechen.

Diese nutzerzentrierte Anpassung gibt den Bürgern das Gefühl der Kontrolle über ihre Mobilitätsentscheidungen. Das Ergebnis ist eine höhere Zufriedenheit, da die Plattform auf persönliche Vorlieben eingeht. Gleichzeitig unterstützt die Lösung städtische Ziele zur Förderung nachhaltiger Mobilität.

6. Fazit

1

Die **wahrgenommene Autonomie** in der Interaktion mit KI-Systemen beeinflusst **Nutzerzufriedenheit** und **Vertrauen**.

2

Automatisierungsstufen und **Anpassungsfähigkeit** haben Einfluss auf die psychologische Wahrnehmung.

3

Transparenz, Flexibilität und **Anpassung der Automatisierungsgrade** stärken die wahrgenommene Autonomie.

4

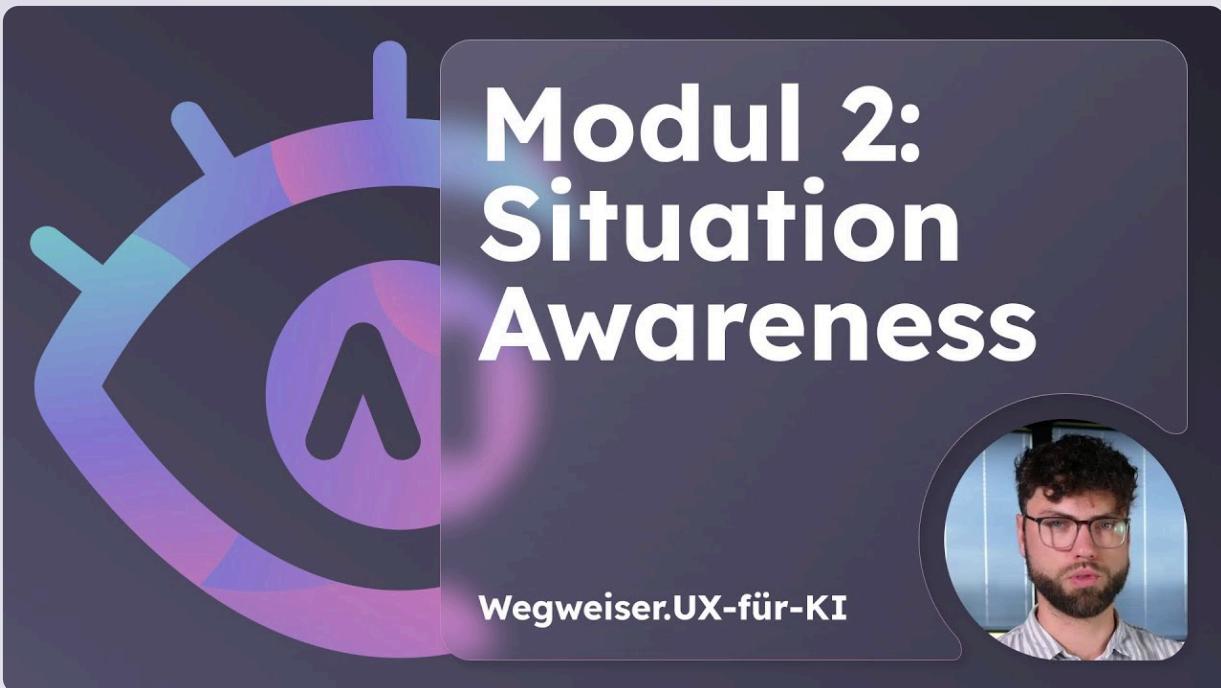
Adaptive Automatisierung und **transparente Entscheidungsprozesse** fördern **Kontrolle** und **Selbstbestimmung**, was zu effektiverer und zufriedenstellender Nutzung führt.

Wahrgenommenes Situationsbewusst- sein

Kursübersicht > KI-bezogene UX

Dies bezieht sich auf das Verständnis der Nutzende über ihre Umgebung und die Änderungen, die durch das KI-System verursacht werden.

Im folgenden Video wird ein Überblick über das wahrgenommene Situationsbewusstsein gegeben.



<https://youtu.be/mhUXTuMQ5mY>

1. Definition **Situationsbewusstsein**

Situationsbewusstsein (Situation Awareness, SA) bedeutet, die Umgebung und ihre wichtigsten Details zu verstehen. Es umfasst auch das Erkennen von Veränderungen, die im Laufe der Zeit oder durch äußere Einflüsse auftreten können. Diese Fähigkeit ist wichtig, um gute Entscheidungen in verschiedenen Situationen zu treffen.

2. Relevante Konzepte und Modelle

Drei Ebenen des Situationsbewusstsein laut Endsley (2000)

- 1. Wahrnehmung:** Die Erfassung der relevanten Informationen in einer Umgebung.
- 2. Verständnis:** Die Interpretation und das Verstehen der Bedeutung dieser Informationen.
- 3. Projektion:** Die Fähigkeit, zukünftige Zustände oder Entwicklungen einer Situation vorherzusagen, basierend auf dem aktuellen Verständnis und der Wahrnehmung der Situation.

Das **HASO-Modell** von Endsley betont ebenfalls diese drei wesentlichen Eigenschaften von Automatisierungssystemen, da sie direkt mit der Situationswahrnehmung der Nutzenden verknüpft sind.

- 1. Transparenz:** Das System muss klar kommunizieren, wie es zu seinen Entscheidungen gelangt.
- 2. Verständlichkeit:** Die Darstellung der Informationen muss leicht nachvollziehbar sein.
- 3. Vorhersehbarkeit:** Nutzende müssen einschätzen können, wie das System sich unter bestimmten Bedingungen verhalten wird.

3. Studien zur User Experience und KI

Studien wie die von **Edgar et al. (2018)** untersuchen das Situationsbewusstsein in der Mensch-Maschine-Interaktion. Sie berechnen das wahrgenommene Situationsbewusstsein auf Basis einer Vertrauensbewertung durch die Bewertung wahr/falsch-Aussagen, wobei hier ebenfalls keine signifikanten Zusammenhang mit dem Verhalten festgestellt werden können. Es ist argumentierbar, dass die Gleichsetzung von wahrgenommenem Situationsbewusstsein und Vertrauen kritisch betrachtet werden kann, da sich das Situationsbewusstsein über drei verschiedene Stufen entwickelt.

Endsley et al. (1995) betont, dass Transparenz, Verständlichkeit und Vorhersehbarkeit entscheidend für die Aufrechterhaltung des Situationsbewusstseins und des Vertrauens in automatisierte Systeme sind. Während Endsley (1998) einerseits betont, dass das wahrgenommene Situationsbewusstsein entscheidend für die Handlungsregulation ist, stellt sie auch fest, dass die Korrelation zwischen wahrgenommenem (oder subjektivem) Situationsbewusstsein oft gering ist.

In der Studie zur User Experience in Digital Contact Tracing (DCT) von **Schrills et al. (2024)** wurde gezeigt, dass das subjektive Situationsbewusstsein der Nutzenden stärker mit der wahrgenommenen Nützlichkeit zusammenhängt als das faktische Situationsbewusstsein.

4. Operationalisierung: Fragebögen und Messinstrumente

Schrills & Franke (2023): SIPA (Subjective Information Processing Awareness)

Die SIPA-Skala ist ein Werkzeug, mit dem man beurteilen kann, wie Erklärungen in der erklärbaren Künstlichen Intelligenz (XAI) auf die Nutzenden wirken. Sie basiert auf den drei Ebenen des Situationsbewusstseins: Transparenz, Verständlichkeit und Vorhersehbarkeit. Mit der SIPA-Skala lässt sich analysieren, wie gut ein System die Nutzenden dabei unterstützt, das Verhalten und die Informationsverarbeitung des Systems nachzuvollziehen.

R. M. Taylor (2017): Situation Awareness Rating Technique (SART)

Die „Situation Awareness Rating Technique“ ist eine Methode zur Bewertung des wahrgenommenen Situationsbewusstseins, die 1990 veröffentlicht wurde. Diese umfasst weitere Konstrukte wie die Arbeitsbelastung und unterscheidet sich daher von Endsleys Konzept des Situationsbewusstseins.

5. Design-Guidelines zur Förderung des Situationsbewusstseins

1. Transparenz sicherstellen

Systeme sollten alle relevanten Elemente der Informationsverarbeitung offenlegen und den Nutzenden zugänglich machen.

Beispiel: Ein KI-gestütztes Dashboard für Ärzte zeigt visuell, welche Daten zur Diagnosestellung verwendet wurden.

2. Verständlichkeit fördern

Intuitive Benutzeroberflächen und kontextbezogene Hilfen sollten die Nutzung von Systemen erleichtern.

Beispiel: Medizinische Diagnose-Tools heben die wichtigsten Informationen hervor und bieten Hilfetexte zur Erklärung komplexer Funktionen.

3. Vorhersehbarkeit verbessern

Systeme sollten Rückmeldungen geben, die die Auswirkungen von Handlungen aufzeigen, z. B. durch Simulationen oder Vorschauen.

Beispiel: Ein System zur Verkehrssteuerung könnte simulieren, wie sich geänderte Ampelphasen auf den Verkehr auswirken, bevor sie tatsächlich implementiert werden.

6. Fazit

1

Situationsbewusstseins (SA) umfasst die **Wahrnehmung**, das **Verständnis** und die **Projektion** relevanter Informationen.

2

Endsley und die SIPA-Facetten betonen die Notwendigkeit von **Transparenz**, **Verständlichkeit** und **Vorhersehbarkeit** für effektives SA.

Wahrgenommene Mentale Belastung

Kursübersicht > KI-bezogene UX

Dieser Aspekt umfasst den kognitiven Aufwand, der erforderlich ist, um Informationen zu verarbeiten und Entscheidungen zu treffen, und die potenzielle Überlastung durch zu viele Informationen.

Was ist die wahrgenommene Mentale Belastung? In diesem Video wird ein Überblick gegeben.



<https://youtu.be/0GpnGeD7BCM>

1. Definition Mentale Belastung

Mentale Arbeitsbelastung (Mental Workload) beschreibt den kognitiven Aufwand, der nötig ist, um eine Aufgabe zu erledigen. Sie umfasst die geistigen Anstrengungen, die notwendig sind, um Informationen zu verarbeiten, Entscheidungen zu treffen und Aktionen durchzuführen.

2. Relevante Konzepte und Modelle

Mentale Arbeitsbelastung ist ein zentrales Konstrukt im Bereich der Automatisierung. Für erklärbare KI (XAI) spielt sie eine besonders wichtige Rolle, da Erklärungen einen paradoxen Effekt haben können: Während KI durch Automatisierung eigentlich eine effizientere Informationsverarbeitung ermöglichen soll, können Erklärungen die mentale Arbeitsbelastung wieder erhöhen. Sie ist daher eine wichtige Metrik für die Bewertung von XAI.

Wenn Menschen durch kognitive Zwänge bewusster über Entscheidungen nachdenken sollen, kann das zwar positive Effekte haben, aber die höhere mentale Anstrengung könnte dazu führen, dass sie diese Methode weniger gerne nutzen - besonders, wenn eine einfachere direkte Empfehlung verfügbar ist.

3. Studien zur User Experience und KI

Ergebnisse mehrerer Studien z. B. von **Sewnath und Crijnen (2021)** und **Tsai et al. (2021)** sowie eine Studie zur Automatisierung von Insulinverabreichungssystemen (AID-Systeme) von **Schrills und Franke (2023)** zeigten, dass der Einsatz von Erklärungen zu einer Informationsüberlastung führen könnte. Zu viele oder zu detaillierte Erklärungen in diesen Systemen beeinträchtigen die Entscheidungsfindung und erhöhen die kognitive Belastung. Insbesondere zeigte sich, dass Systeme mit hoher Informationsoffnenlegung nicht immer zu besseren Ergebnissen führten, sondern manchmal genau das Gegenteil bewirken.

4. Operationalisierung: Fragebögen und Messinstrumente

Hart (2006): NASA Task Load Index (NASA-TLX)

Eine weit verbreitete Methode zur Messung der mentale Arbeitsbelastung ist der **NASA Task Load Index (NASA-TLX)**. Dieser Fragebogen bewertet verschiedene Dimensionen der Arbeitsbelastung, darunter:

- Mentale Anforderungen
- Physische Anforderungen
- Zeitliche Anforderungen
- Leistung
- Anstrengung
- Frustration

Der NASA-TLX ist eine hilfreiche Methode, um die kognitive Belastung in verschiedenen Arbeitsumgebungen, einschließlich der Nutzung von KI-Systemen, zu erfassen.

5. Design-Guidelines zur Reduktion der kognitiven Belastung

1. Vereinfachung von Informationen

Informationen sollten minimalistisch dargestellt und nur schrittweise offengelegt werden, um die kognitive Belastung zu reduzieren.

Beispiel: Ein KI-gestütztes Dashboard zeigt nur die wesentlichen Informationen und blendet zusätzliche Details bei Bedarf ein.

2. Anpassbarkeit der Informationsmenge

Systeme sollten es den Nutzenden ermöglichen, die Menge an Informationen anzupassen, die sie benötigen, um Entscheidungen zu treffen.

Beispiel: Personalisierte Einstellungen, die es den Nutzenden erlauben, festzulegen, wie viele Details sie sehen möchten.

3. Fokussierung auf kritische Informationen

Das System sollte kontextbezogen die wichtigsten Informationen priorisieren.

Beispiel: In einem Verkehrsleitsystem werden in einer Notfallsituation nur die kritischsten Daten hervorgehoben, wie z. B. gesperrte Straßen oder gefährliche Wetterbedingungen.

4. Reduktion der kognitiven Belastung durch Automatisierung

Routinetätigkeiten sollten automatisiert werden, um die geistige Anstrengung der Nutzenden zu minimieren.

Beispiel: Ein KI-System im Verwaltungsbereich könnte die automatische Überprüfung von Anträgen übernehmen, sodass sich Mitarbeitende auf Ausnahmen und komplexe Fälle konzentrieren können.

6. Fazit

1

Mentale Arbeitsbelastung ist in der **erklärbaren KI (XAI)** entscheidend.

2

Automatisierung erleichtert die Verarbeitung, **Erklärungen** können jedoch **Informationsüberlastung** verursachen.

3

Wichtige Systemmerkmale: **Reduzierung, Anpassbarkeit und Priorisierung von Informationen**, um Nutzer effektiv zu entlasten.

4

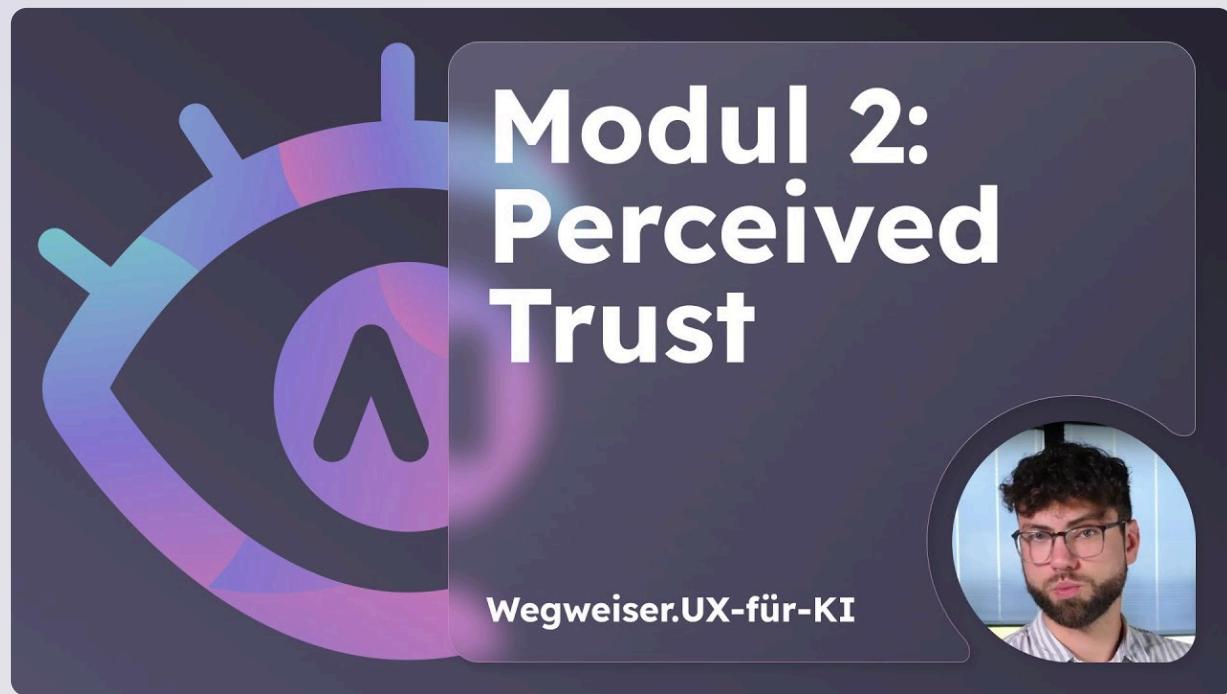
Der **NASA-TLX** hilft bei der Bewertung der Arbeitsbelastung und Verbesserung der Benutzerfreundlichkeit.

Wahrgenommene Vertrauenswürdig- keit

Kursübersicht > KI-bezogene UX

Damit ist das Vertrauen gemeint, das Nutzende in ein KI-System haben, basierend auf dessen Handlungen.

Dieses Video zeigt wie man einem System vertrauen kann und welche Formen es gibt.



1. Definition Wahrgenommene Vertrauenswürdigkeit

Im Kontext von KI-Systemen beschreibt die wahrgenommene Vertrauenswürdigkeit (perceived trustworthiness) das Vertrauen, das Nutzende einem KI-System entgegenbringen, basierend auf ihrer Wahrnehmung von dessen Zuverlässigkeit, Verständlichkeit und ethischen Prinzipien. Vertrauen wird als die Bereitschaft definiert, sich auf die Entscheidungen eines Systems zu verlassen, ohne es direkt überwachen oder kontrollieren zu können.

Vertrauen ist zentral für die Akzeptanz und Nutzung von KI-Systemen, insbesondere im Bereich der erklärbaren Künstlichen Intelligenz (XAI), wo ethische und soziale Aspekte besonders relevant sind.

2. Relevante Konzepte und Modelle

Drei Schlüsselfaktoren des Vertrauens in Systeme nach Mayer et al. (1995):

- 1. Fähigkeit:** Die Kompetenz und Fertigkeiten des Systems, seine Aufgaben korrekt auszuführen.
- 2. Wohlwollen:** Das Maß, in dem das System im besten Interesse des Nutzenden handelt.
- 3. Integrität:** Das Vertrauen, dass das System nach ethischen Prinzipien und Regeln handelt.

Vertrauen als eine Einstellung nach Lee & See (2004)

Vertrauen wird als die Einstellung beschrieben, dass ein Agent/System dazu beitragen wird, die Ziele einer Person in einer Situation zu erreichen, die durch Unsicherheit und Verwundbarkeit gekennzeichnet ist. Es ist wichtig zu beachten, dass Vertrauen als eine Einstellung und nicht als ein Verhalten konzeptualisiert werden sollte. Dabei sind drei Faktoren entscheidend:

- 1. Leistung:** Die Genauigkeit und Zuverlässigkeit des Systems.
- 2. Zweck:** Die Ziele und Absichten, die das System verfolgt.
- 3. Prozess:** Die Methoden und Verfahren, die das System verwendet.

Kognitives vs. affektives Vertrauen nach Madsen und Gregor (2000):

Sie beschreiben Vertrauen als „das Ausmaß, in dem ein Nutzer Vertrauen in die Empfehlungen, Handlungen und Entscheidungen einer künstlichen Entscheidungsunterstützung hat und bereit ist, auf deren Basis zu handeln.“

Sie unterscheiden zwischen:

- 1. Kognitives Vertrauen:** Basierend auf den wahrgenommenen Eigenschaften des Systems, wie Zuverlässigkeit und Verständlichkeit. Wenn das System beispielsweise transparent ist und nachvollziehbare Entscheidungen trifft, steigt das Vertrauen.
- 2. Affektives Vertrauen:** Emotionale Bindungen oder persönliche Erfahrungen mit dem System fördern das Vertrauen, besonders bei sprachbasierten Assistenzsystemen oder Robotern.

Drei Dimensionen des Vertrauens in der Automatisierung nach Hoff und Bashir (2015):

- 1. Dispositionelles Vertrauen:** bezieht sich auf die generelle Tendenz eines Nutzers, Automatisierung zu vertrauen, basierend auf Persönlichkeit und bisherigen Erfahrungen.
- 2. Situatives Vertrauen:** wird durch den spezifischen Kontext beeinflusst, in dem die Automatisierung verwendet wird, einschließlich Aufgabenmerkmalen und Umweltfaktoren.
- 3. Erlerntes Vertrauen:** entwickelt sich über die Zeit durch Interaktionen mit dem System, wobei positive Erfahrungen das Vertrauen stärken und negative es schwächen.

Dieses Modell ist besonders wertvoll, da es die dynamische Natur von Vertrauen und die Bedeutung der Nutzererfahrungen im Zeitverlauf hervorhebt.

Vertrauenswürdigkeits-Hinweisen nach Schlicker et al. (2022):

Vertrauenswürdigkeits-Hinweise sind entscheidend für die wahrgenommene Vertrauenswürdigkeit eines Systems. Hinweise zur Informationsverarbeitung zeigen nicht nur, wie zuverlässig das System ist, sondern helfen den Nutzern auch, dessen Funktionsweise besser zu verstehen und seine Fähigkeit zur Aufgabenbewältigung einzuschätzen. Hinweise, die nicht mit den Merkmalen der Informationsverarbeitung des Systems zusammenhängen, z. B. die Reputation des Herstellers oder soziale Hinweise wie das Nutzungsverhalten von Personen im eigenen Umfeld können ebenfalls relevant sein.

3. Studien zur User Experience und KI

Schrills (2024): Einfluss der Systemzuverlässigkeit auf Nutzervertrauen in KI-Systeme

Die Studie zeigt, dass die Selbsteinschätzung des Vertrauens in ein KI-System keinen direkten Einfluss auf das tatsächliche Vertrauen in dessen Empfehlungen hat. Stattdessen erwies sich die angegebene Zuverlässigkeit des Systems als der stärkste Einflussfaktor auf das Verhalten der Nutzenden. Das bedeutet, dass Nutzende ihr Vertrauen eher auf die wahrgenommene Zuverlässigkeit des Systems stützen als auf ihr eigenes, subjektives Vertrauensempfinden. Dieses Ergebnis legt nahe, dass die Kommunikation und Darstellung der Zuverlässigkeit eines KI-Systems einen größeren Einfluss auf die Nutzung hat als traditionelle Methoden zur Messung von Vertrauen.

Vereschak et al. (2024): Menschliche Einflüsse und Stakeholder-spezifische Anforderungen

- 1. Vertrauensanforderungen:** Die Teilnehmenden identifizierten wesentliche Elemente für Vertrauen und unterschieden es von Konzepten wie Vertrauenswürdigkeit, Verlass und Befolgung. Positive Erwartungen und wahrgenommenes Risiko waren dabei entscheidende Faktoren, wobei die Komplexität der Aufgabe als zusätzliche Voraussetzung für Vertrauen in KI-gestützte Entscheidungsprozesse hervorgehoben wurde.
- 2. Menschlicher Einfluss auf Vertrauen:** Vertrauen in KI-Systeme wurde stark von menschlichen Akteuren beeinflusst, etwa von den Personen, die das System entwickeln und einsetzen, und nicht nur von den technischen Merkmalen des Systems.
- 3. Stakeholderspezifische Vertrauensfaktoren:** Die Faktoren, die das Vertrauen zwischen Mensch und KI beeinflussen, variieren je nach Stakeholder. Zum Beispiel legen Entscheidungsträger, die das System nutzen, und diejenigen, die von den Entscheidungen betroffen sind (z. B. Patienten im medizinischen Kontext), auf unterschiedliche Vertrauensaspekte Wert.

4. Operationalisierung: Fragebögen und Messinstrumente

Jian et al. (2001): Trust in Automation (TiA)-Skala

Die Skala ist eine der am häufigsten verwendeten Skalen zur Selbsteinschätzung von Vertrauen. Sie bewertet, wie sehr Nutzer die Leistungsfähigkeit und Zuverlässigkeit automatisierter Systeme einschätzen und unterscheidet dabei zwischen Mensch-Mensch- und Mensch-Maschine-Vertrauen. Die Skala erfasst zentrale Vertrauensfaktoren und hilft, das Vertrauen der Nutzer in Automation und dessen Einfluss auf die Nutzung zu verstehen.

5. Design-Guidelines zur Förderung der Vertrauenswürdigkeit

1. Förderung von Transparenz und Verständlichkeit

Erklärbare KI (XAI) ermöglicht, dass die Entscheidungen des Systems klar und nachvollziehbar sind, was das Vertrauen der Nutzer stärkt.

Beispiel: Ein Finanzplanungssystem, das seine Berechnungen offenlegt und erklärt, welche Parameter berücksichtigt wurden, hilft Nutzern, die Entscheidungsprozesse nachzuvollziehen.

2. Sicherstellung von Systemleistung und Zuverlässigkeit

Systeme sollten verlässlich arbeiten und klare Rückmeldungen geben, um Vertrauen aufzubauen und Unsicherheiten zu reduzieren.

Beispiel: Ein medizinisches KI-System, das bei der Diagnose Fehlerindikatoren und Bestätigungen anzeigt, schafft Vertrauen in die Genauigkeit und Zuverlässigkeit der Ergebnisse.

3. Stärkung des erlernten Vertrauens

Einführungen und Schulungen können dispositionelles Vertrauen fördern, indem sie das Verständnis und die Vertrautheit mit dem System verbessern.

Beispiel: Nutzerfreundliche Tutorials für ein neues KI-gestütztes Verwaltungssystem fördern das grundlegende Vertrauen der Nutzer in das System.

4. Förderung von affektivem Vertrauen

Eine personalisierte Nutzererfahrung und positive User Experience (UX) tragen zur Entwicklung von affektivem Vertrauen bei.

Beispiel: Ein Sprachassistentensystem, das personalisierte Präferenzen berücksichtigt, stärkt die emotionale Bindung und das Vertrauen der Nutzer in das System.

5. Kontextbezogene Erklärungen für Empfehlungen geben

Um situatives Vertrauen zu fördern, sollten Systeme kontextabhängige Erklärungen liefern, die auf die spezifische Entscheidungssituation der Nutzer eingehen.

Beispiel: Ein Personalplanungssystem für das Gesundheitswesen erklärt, warum bestimmte Dienstpläne vorgeschlagen werden, z. B. basierend auf Arbeitslasten oder Personalverfügbarkeit.

6. Fazit

1

Vertrauen ist entscheidend: Vertrauenswürdigkeit fördert die Akzeptanz von KI-Systemen.

2

Unterschiedliche Vertrauensfaktoren: Fähigkeit, Integrität und Wohlwollen sind grundlegende Faktoren für Vertrauen.

3

Menschliche Einflüsse berücksichtigen: Vertrauen wird nicht nur durch das System, sondern auch durch menschliche Interaktionen und den sozialen Kontext beeinflusst.

4

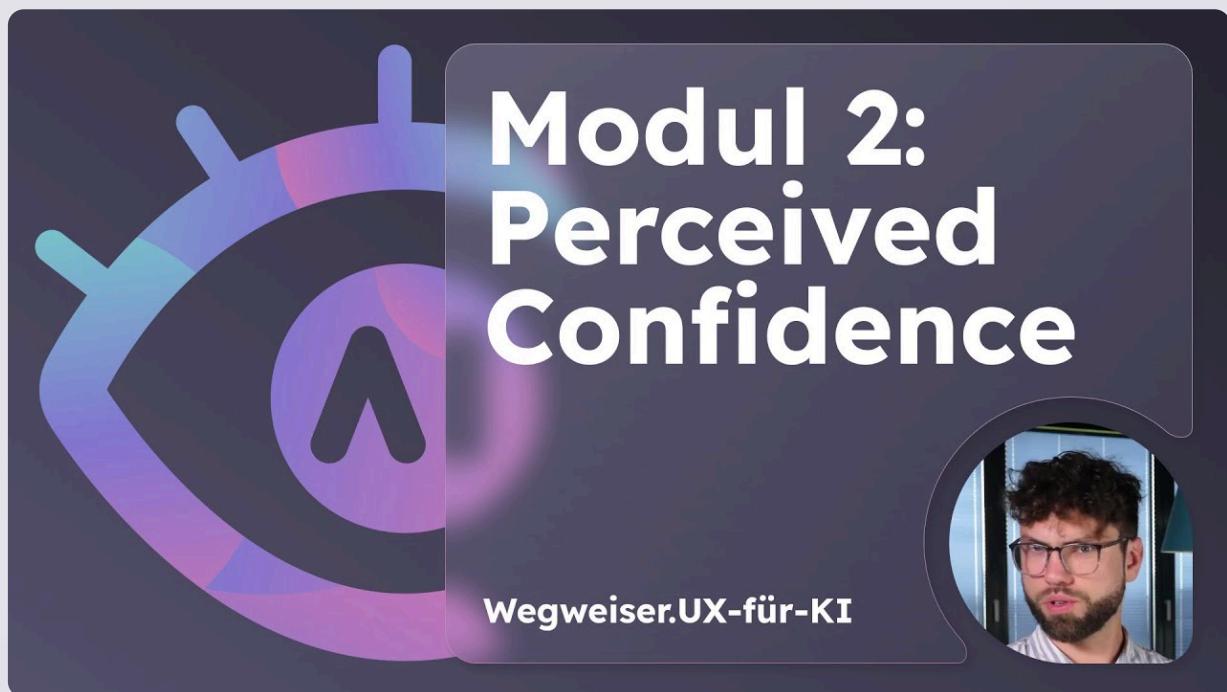
Zuverlässigkeit sichtbar machen: Nutzervertrauen basiert stark auf der wahrgenommenen Zuverlässigkeit des Systems.

Wahrgenommene Diagnostizität

Kursübersicht > KI-bezogene UX

Dies beschreibt das Vertrauen der Nutzende in die Diagnosen oder Vorschläge des KI-Systems und wie gut diese die gewünschten Ergebnisse liefern.

Im folgenden Video wird ein Überblick über den Begriff Diagnostizität gegeben.



<https://youtu.be/jdsnWIIGT7c>

1. Definition wahrgenommene Diagnostizität

Die wahrgenommene Diagnostizität (Perceived Confidence/Diagnosticity) beschreibt, wie nützlich ein System für Nutzende ist, um fundierte Entscheidungen zu treffen. Ein System mit hoher Diagnostizität unterstützt die Bewertung verschiedener Optionen, sodass Nutzende auf Basis der vom System bereitgestellten Informationen die beste Entscheidung treffen können.

Je präziser und hilfreicher die Hinweise sind, desto mehr steigt die diagnostische Qualität des Systems und stärkt damit das Vertrauen der Nutzenden.

2. Relevante Konzepte und Modelle

Auswahl von Handlungsmöglichkeiten

Aus psychologischer Sicht ist Diagnostizität entscheidend, weil sie die Auswahl einer Handlung erleichtert. Informationen mit hoher Diagnostizität unterstützen Menschen dabei, gezielte Entscheidungen zu treffen. So kann ein diagnostisches System im Gesundheitswesen dazu beitragen, zwischen wenigen klaren Hypothesen zu unterscheiden, was letztlich Unsicherheit reduziert und die Entscheidungsfindung erleichtert.

Ein Beispiel dafür ist ein KI-System im Gesundheitswesen, das bei Symptomen wie Fieber die Wahrscheinlichkeiten für mögliche Erkrankungen wie Grippe oder Erkältung analysiert. Durch das Bereitstellen spezifischer Informationen hilft das System, die Entscheidung über den nächsten Schritt gezielt zu erleichtern.

Unterschied zwischen wahrgenommener und tatsächlicher Diagnostizität

Manchmal schätzen Nutzende bestimmte Informationen als sehr hilfreich für eine Entscheidung ein, obwohl diese Infos gar nicht wirklich dabei helfen, zwischen verschiedenen Möglichkeiten zu unterscheiden. Dies kann dazu führen, dass unwichtige Details überbewertet werden und dadurch Fehlentscheidungen entstehen (Nelson, 2005)

Beispielsweise hat das Symptom „Fieber“ bei der Unterscheidung zwischen COVID-19 und Grippe nur einen geringen diagnostischen Wert, auch wenn es allgemein zur Krankheitsdiagnose beiträgt.

Dieser Effekt führt zu sogenannten „Pseudo-Diagnosen“, bei denen Menschen diagnostisch wenig hilfreiche Daten auswählen und diese für ihre Entscheidung nutzen (Kern & Doherty, 1982). Ähnlich wie bei der „Illusion of Explanatory Depth“ (Chromik et al., 2021) können dadurch Fehleinschätzungen der eigenen Leistung und eine falsche Bewertung des Systems entstehen.

Unterschied zwischen Informationswert und Diagnostizität

Der Begriff „Informationswert“ bezieht sich darauf, wie stark eine Information generell Unsicherheit verringert. „Diagnostizität“ hingegen ist spezifischer: Sie hilft, zwischen bestimmten Möglichkeiten (z. B. Grippe vs. Erkältung) zu unterscheiden. Wenn es nur wenige klare Optionen gibt, ist Diagnostizität wichtiger. Wenn es viele mögliche Optionen gibt, hilft der allgemeine Informationswert mehr.

3. Studien zur User Experience und KI

Schrills und Franke (2023): Einfluss der Diagnostizität auf Vertrauen und Nutzungsabsicht

Schrills und Franke (2023) betonen, dass die wahrgenommene Diagnostizität entscheidend für die Vertrauenswürdigkeit und Nutzung von Systemen ist. Bei Digital Contact Tracing (DCT)-Apps zur Pandemiekontaktverfolgung bevorzugen Nutzende detaillierte und klare Informationen, besonders diagnostische Details wie das Tragen einer Maske während einer Pandemie. Dies stärkt ihr Vertrauen in die Genaugigkeit und Nützlichkeit der App und erhöht somit auch Zufriedenheit. Das Fehlen diagnostischer Hinweise beeinflusst das Verhalten der Nutzende und bringt sie dazu, eher allgemeine, weniger gezielte Informationen zu nutzen.

Bartlett und McCarley (2017): Suboptimale Entscheidungen durch fehlende Diagnostizität

Wenn eine KI keine spezifisch diagnostischen Informationen liefert, neigen Nutzenden oft dazu, alternative Strategien anzuwenden, die nicht immer optimal sind. Die Studie beschreibt, dass Menschen in solchen Situationen oft eine Strategie namens „probability matching“ verwenden. Dabei passen sie ihre Entscheidungen an die allgemeine Zuverlässigkeit der KI an, anstatt auf gezielte, diagnostische Hinweise zu achten.

4. Operationalisierung: Fragebögen und Messinstrumente

Wahrgenommene Ergebnisdagnostik

Hier wird erfasst, inwieweit die Nutzer der Meinung sind, dass die bereitgestellten Informationen fundiert sind und zu einer besseren Entscheidungsfindung beitragen.

Fragebögen zur Messung von Vertrauenswürdigkeit und Diagnostizität:

Diese Instrumente bewerten, wie präzise und klar die Informationen eines Systems wahrgenommen werden und wie stark sie das Vertrauen der Nutzer beeinflussen.

5. Design-Guidelines

Um Nutzer das Verständnis und Vertrauen in die Entscheidungen einer Künstlichen Intelligenz (KI) zu erleichtern, gibt es verschiedene Techniken in der Erklärbaren KI (XAI). Diese Techniken helfen, Entscheidungen der KI verständlicher und transparenter zu machen.

1. Kontrafaktische Erklärungen – „Was wäre wenn“-Szenarien

Kontrafaktische Erklärungen zeigen, wie eine kleine Änderung an den Eingabedaten zu einer anderen Entscheidung der KI führen könnte. Zum Beispiel: Wenn eine Kreditbewilligung abgelehnt wird, könnte die KI erklären, dass eine Erhöhung des Einkommens um einen bestimmten Betrag zur Bewilligung geführt hätte. Solche „Was wäre wenn“-Erklärungen helfen den Nutzenden, die Entscheidungsgrenzen der KI zu verstehen (Warren et al., 2022).

2. Semantische Anreicherung und Heatmaps – Verständlichere Visualisierungen

Durch semantische Anreicherung, also durch das Hinzufügen von zusätzlichen Erklärungen, werden Heatmaps (visuelle Darstellungen) verständlicher. Zum Beispiel in der medizinischen Bildgebung können Heatmaps aufzeigen, welche Bildbereiche für eine Diagnose besonders wichtig waren. Diese zusätzlichen Details helfen den Nutzenden, die Entscheidungslogik der KI besser nachzuvollziehen und das Vertrauen zu stärken (Gianfagna & Di Cecco, 2021; Tonekaboni et al., 2019).

3. Konfidenzbewertungen – Vertrauen durch Unsicherheitsangaben

Konfidenzbewertungen zeigen, wie sicher oder unsicher die KI bei ihren Vorhersagen ist. Diese Informationen sind besonders wichtig in Bereichen wie Gesundheit und Finanzen, wo Entscheidungen große Auswirkungen haben können. Wenn die KI ihre Unsicherheit angibt, können Nutzende besser einschätzen, ob sie der Entscheidung vertrauen möchten oder nicht (Gianfagna and Di Cecco, 2021; T. Le et al., 2023).

4. Lokale Feature-Relevanz – Bedeutung einzelner Merkmale

Die lokale Feature-Relevanz zeigt an, welche spezifischen Eingabemerkmale zu einer bestimmten Entscheidung der KI geführt haben. Beispielsweise könnte in einem System zur Betrugserkennung hervorgehoben werden, dass ungewöhnliche Transaktionsbeträge oder -orte eine große Rolle gespielt haben. Diese Detailinformationen helfen den Nutzenden, die Entscheidungen der KI mit ihrem eigenen Wissen zu vergleichen und zu validieren (Doshi-Velez and Kim, 2017; Lundberg et al., 2019).

6. Fazit

1

Die wahrgenommene Diagnostizität eines KI-Systems ist entscheidend für fundierte und präzise Entscheidungen.

2

Systeme, die klare diagnostische Informationen bereitstellen, stärken das Vertrauen der Nutzenden sowie die Akzeptanz und effektive Nutzung der Technologie.

3

Besonders wichtig ist dies in kritischen Bereichen wie dem Gesundheitswesen.

4

Die diagnostische Qualität eines Systems hat direkten Einfluss auf das Wohl der Gemeinschaft.

Zusammenfassung und Ausblick

Kursübersicht > KI-bezogene UX

Zum Ende des Moduls wird eine kurze Zusammenfassung über die Inhalte des Moduls gegeben.



<https://youtu.be/Qmg1FIFCSMw>

In dieser Lektion haben wir uns eingehend mit den spezifischen Aspekten der User Experience (UX) auseinandergesetzt, die im Kontext von KI-Systemen eine besondere Rolle spielen.

Die 5 UX-Kernaspekte und deren Designrichtlinien

1. Adaptive Automatisierung für wahrgenommene Autonomie

Wie sehr fühlen sich Nutzende in der Lage, selbstständig Entscheidungen zu treffen und zu handeln, während sie mit einem KI-System interagieren? Nutzende sollten die Kontrolle darüber haben, wie stark sie das System automatisieren oder manuell bedienen möchten. Dies fördert das Gefühl der Autonomie und gibt ihnen die Flexibilität, sich bei Bedarf stärker auf ihre eigenen Entscheidungen zu verlassen.

Bieten Sie Nutzenden die Möglichkeit, zwischen verschiedenen Automatisierungsstufen zu wechseln, sodass sie je nach Präferenz oder Aufgabenanforderung selbst entscheiden können, wie viel Kontrolle sie dem System überlassen.

2. Transparenz zur Unterstützung des Situationsbewusstseins

Situationsbewusstsein ist das Verständnis und die Wahrnehmung der Nutzenden über die aktuelle Umgebung und die Auswirkungen der KI auf diese. Ein hohes Maß an Situationsbewusstsein erfordert, dass Nutzende jederzeit relevante Informationen über das System und dessen Entscheidungen erhalten. Transparenz fördert das Verständnis für die Funktionsweise und Entscheidungsgrundlagen des Systems.

Stellen Sie sicher, dass das System die Datenquellen, Prozesse und Faktoren, die eine Entscheidung beeinflussen, klar kommuniziert. Nutzen Sie visuelle Darstellungen oder erklärende Hinweise, um komplexe Abläufe verständlicher zu machen.

3. Flexibilität bei der Informationsverarbeitung zur Reduzierung der mentalen Belastung

Mentale Belastung beschreibt den kognitiven Aufwand, der durch die Verarbeitung von Informationen entsteht und die potenzielle Überlastung durch zu viele Daten. Da zu viele oder unstrukturierte Informationen zu kognitiver Überlastung führen können, ist es entscheidend, Nutzende die Kontrolle über die Art und Menge der angezeigten Informationen zu geben. Dies hilft, die wahrgenommene mentale Belastung zu reduzieren.

Integrieren Sie Funktionen, die es Nutzende ermöglichen, die Anzeige von Informationen nach Bedarf zu filtern, zu kategorisieren oder zu priorisieren. Zum Beispiel kann eine Zusammenfassungsansicht für weniger erfahrene Nutzer und eine Detailansicht für Experten angeboten werden.

4. Zuverlässigkeit und klare Darstellung zur Förderung der wahrgenommenen Vertrauenswürdigkeit

Vertrauen entsteht, wenn das System nicht nur zuverlässig und präzise arbeitet, sondern diese Eigenschaften auch klar vermittelt. Nutzende müssen darauf vertrauen können, dass das System korrekt und ethisch agiert.

Verwenden Sie visuelle Indikatoren, die die Zuverlässigkeit und Erfolgsquote des Systems darstellen, und bieten Sie Erklärungen, die das ethische und technische Verhalten des Systems untermauern.

5. Nutzerzentrierte Anpassung zur Unterstützung der wahrgenommenen Diagnostizität

Diagnostizität ist die Fähigkeit eines Systems, Nutzende mit präzisen Informationen bei fundierten Entscheidungen zu unterstützen. Nutzende sollten das System an ihre individuellen Präferenzen und Informationsbedürfnisse anpassen können, damit es sie optimal bei der Entscheidungsfindung unterstützt. Dies stärkt die Fähigkeit des Systems, genaue und hilfreiche Informationen bereitzustellen.

Ermöglichen Sie personalisierte Einstellungen, durch die Nutzenden entscheiden können, welche Art von Informationen angezeigt werden und wie detailliert diese sein sollen. Ein personalisiertes Dashboard oder konfigurierbare Berichte können helfen, das System effizienter zu nutzen.

Ausblick auf das Modul: "Gestaltungsziele für menschzentrierte KI"

Im kommenden Modul werden wir uns mit spezifischen Eigenschaften von KI-Systemen beschäftigen, die besonders wichtig für die UX sind: Erklärbarkeit, Nachvollziehbarkeit, Vertrauenswürdigkeit, Kontrollierbarkeit und Transparenz. Wir werden untersuchen, wie diese Eigenschaften die Nutzererfahrung beeinflussen und welche Prinzipien bei der Gestaltung von KI-Systemen berücksichtigt werden sollten, um eine positive und vertrauensvolle Interaktion zu gewährleisten.

Wir hoffen, dass diese Lektion Ihnen wertvolle Einblicke in die Relevanz von UX-Konstrukten bei der Gestaltung von KI-Systemen gegeben hat. Diese Konzepte sind nicht nur für die Entwicklung von KI-Systemen entscheidend, sondern auch für die Bewertung, wie gut diese Systeme die Bedürfnisse und Erwartungen der Nutzer erfüllen. Bereiten Sie sich darauf vor, tiefer in die Dimensionen einzutauchen, die das Nutzererlebnis mit KI-Systemen weiter prägen werden.

08

Quellen

Kursübersicht > KI-bezogene UX

Literatur zu Wahrgenommener Autonomie

- Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. **IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans**, 30(3), 286-297. <https://doi.org/10.1109/3468.844354>
- Hopkins, D., & Schwanen, T. (2021). Talking about automated vehicles: What do levels of automation do? **Technology in Society**, 64, 101488. <https://doi.org/10.1016/j.techsoc.2020.101488>
- Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. **Human Factors: The Journal of the Human Factors and Ergonomics Society**, 56(3), 476-488. <https://doi.org/10.1177/0018720813501549>
- Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. **Theoretical Issues in Ergonomics Science**, 5(2), 113-153. <https://doi.org/10.1080/1463922021000054335>

- Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports*, 12, 3768. <https://doi.org/10.1038/s41598-022-07808-x>
- Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227–268.
- Zoubir, M. (2024). **Preference for Automation Types Scale (PATS)**. <https://doi.org/10.13140/RG.2.2.22149.97769>
- Moradbakhti, L., Leichtmann, B., & Mara, M. (2024). Development and validation of a basic psychological needs scale for technology use. *Psychological Test Adaptation and Development*, 5(1), 26–45. <https://doi.org/10.1027/2698-1866/a000062>

Literatur zu Wahrgenommenem Situationsbewusstsein (SA)

- Edgar, G. K., Catherwood, D., Baker, S., Sallis, G., Bertels, M., Edgar, H. E., Nikolla, D., Buckle, S., Goodwin, C. & Whelan, A. (2018). Quantitative Analysis of Situation Awareness (QASA): Modelling and measuring situation awareness using signal detection theory. *Ergonomics*, 61(6), 762–777. <https://doi.org/10.1080/00140139.2017.1420238>
- Endsley, M., Sollenberger, R. & Stein, E. (2000). Situation awareness: A comparison of measures. In **Proceedings of the Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium**, Savannah, GA.

- Endsley, M. R. (1995). Toward a Theory of Situation Awareness in Dynamic Systems. **Human Factors**, 37(1), 32–64.
<https://doi.org/10.1518/001872095779049543>
- Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned From Human-Automation Research. **Human Factors: The Journal of the Human Factors and Ergonomics Society**, 59(1), 5–27.
<https://doi.org/10.1177/0018720816681350>
- Endsley, M. R., Selcon, S. J., Hardiman, T. D. & Croft, D. G. (1998). A Comparative Analysis of Sagat and Sart for Evaluations of Situation Awareness. In **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, 42(1), 82–86.
<https://doi.org/10.1177/154193129804200119>
- Schrills, T., & Franke, T. (2023). Wie erleben Nutzer die Nachvollziehbarkeit von KI-Systemen? Untersuchung des subjektiven Informationsverarbeitungsbewusstseins in automatisierten Insulinabgabesystemen (AID-Systemen). *ACM Transactions on Interactive Intelligent Systems*, 13(4), 25:1–25:34.
<https://doi.org/10.1145/3588594>
- Schrills, T., Kojan, L., Gruner, M., Calero Valdez, A. & Franke, T. (2024). Effects of User Experience in Automated Information Processing on Perceived Usefulness of Digital Contact-Tracing Apps: Cross-Sectional Survey Study. **JMIR Human Factors**, 11, e53940.
<https://doi.org/10.2196/53940>
- Taylor, R. M. (2017). Situational Awareness Rating Technique (Sart): The Development of a Tool for Aircrew Systems Design. In E. Salas (Hrsg.), **Situational Awareness** (1. Aufl., S. 111–128). Routledge.
<https://doi.org/10.4324/9781315087924-8>

Literatur zu Wahrgenommene Mentale Arbeitsbelastung

- Miller, T. (2023). Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI. **2023 ACM Conference on Fairness, Accountability, and Transparency**, 333–342.
<https://doi.org/10.1145/3593013.3594001>
- Vidulich, M. A., & Tsang, P. S. (2012). Mental Workload and Situation Awareness. In G. Salvendy (Hrsg.), **Handbook of Human Factors and Ergonomics** (1. Aufl., S. 243–273). Wiley.
<https://doi.org/10.1002/9781118131350.ch8>
- Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human Mental Workload: A Survey and a Novel Inclusive Definition. **Frontiers in Psychology**, 13, 883321.
<https://doi.org/10.3389/fpsyg.2022.883321>
- Sewnath, G., & Crijnen, J. (2021). **How much is too much? Levels of AI Explainability within Decision Support Systems' User Interfaces for improved decision-making performance.**
- Tsai, C.-H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**, 1–17.
<https://doi.org/10.1145/3411764.3445101>
- Schrills, T., & Franke, T. (2023). How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. **ACM Transactions on Interactive Intelligent Systems**, 13(4), 1–34.
<https://doi.org/10.1145/3588594>

- Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, 50 (9), 904–908.
<https://doi.org/10.1177/154193120605000909>

Literatur zu Wahrgenommene Vertrauenswürdigkeit

- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. **The Academy of Management Review**, 20(3), 709. <https://doi.org/10.2307/258792>
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. **Human Factors: The Journal of the Human Factors and Ergonomics Society** , 46(1), 50–80.
https://doi.org/10.1518/hfes.46.1.50_30392
- Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. **11th australasian conference on information systems** , 53, 6–8.
- Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. **Human Factors: The Journal of the Human Factors and Ergonomics Society** , 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M. C., & Langer, M. (2022). **How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM)**.<https://doi.org/10.31234/osf.io/qhwvx>

- Schrills, T., & Franke, T. (2023). How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. **ACM Transactions on Interactive Intelligent Systems**, 13(4), 1–34.
<https://doi.org/10.1145/3588594>
- Vereschak, O., Alizadeh, F., Bailly, G., & Caramiaux, B. (2024). Trust in AI-assisted Decision Making: Perspectives from Those Behind the System and Those for Whom the Decision is Made. **Proceedings of the CHI Conference on Human Factors in Computing Systems**, 1–14.<https://doi.org/10.1145/3613904.3642018>
- Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. **Frontiers in psychology**, 12, 604977.
<https://doi.org/10.3389/fpsyg.2021.604977>

Literatur zu Wahrgenommene Confidence / Diagnosticity

- Wickens, C. D., & Scott, B. D. (1983). A comparison of verbal and graphical information presentation in a complex information integration decision task. In Tech. Rep. EPL-83-1/ONR-83-1. Engineering-Psychology Research Laboratory, University of Illinois Urbana.
- Nelson, J. D. (2005). Finding Useful Questions: On Bayesian Diagnosticity, Probability, Impact, and Information Gain. **Psychological Review**, 112(4), 979–999.
<https://doi.org/10.1037/0033-295X.112.4.979>

- Kern, L., & Doherty, M. E. (1982). ‘Pseudodiagnosticity’ in an idealized medical problem-solving environment. **Academic Medicine**, **57**(2), 100–104. <https://doi.org/10.1097/00001888-198202000-00004>
- Chromik, M., Eiband, M., Buchner, F., Krüger, A., & Butz, A. (2021). I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. **26th International Conference on Intelligent User Interfaces**, 307–317. <https://doi.org/10.1145/3397481.3450644>
- Schrills, T., & Franke, T. (2023). How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. **ACM Transactions on Interactive Intelligent Systems**, **13**(4), 1–34. <https://doi.org/10.1145/3588594>
- Bartlett, M. L., & McCarley, J. S. (2017). Benchmarking Aided Decision Making in a Signal Detection Task. **Human Factors: The Journal of the Human Factors and Ergonomics Society**, **59**(6), 881–900. <https://doi.org/10.1177/0018720817700258>
- Warren, G., Smyth, B., & Keane, M. T. (2022). “Better” Counterfactuals, Ones People Can Understand: Psychologically-Plausible Case-Based Counterfactuals Using Categorical Features for Explainable AI (XAI). In M. T. Keane & N. Wiratunga (Eds.), **Case-Based Reasoning Research and Development** (Vol. 13405, pp. 63–78). Springer International Publishing. https://doi.org/10.1007/978-3-031-14923-8_5
- Gianfagna, L., & Di Cecco, A. (2021). Explainable AI: Needs, Opportunities, and Challenges. In L. Gianfagna & A. Di Cecco, **Explainable AI with Python** (pp. 27–46). Springer International Publishing. https://doi.org/10.1007/978-3-030-68640-6_2

- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), **Proceedings of the 4th Machine Learning for Healthcare Conference** (Vol. 106, pp. 359–380). PMLR.

<https://proceedings.mlr.press/v106/tonekaboni19a.html>

- Le, T., Miller, T., Singh, R., & Sonenberg, L. (2023). Explaining Model Confidence Using Counterfactuals. **Proceedings of the AAAI Conference on Artificial Intelligence**, 37(10), 11856–11864.

<https://doi.org/10.1609/aaai.v37i10.26399>

- Doshi-Velez, F., & Kim, B. (2017). **Towards A Rigorous Science of Interpretable Machine Learning** (No. arXiv:1702.08608). arXiv.

<http://arxiv.org/abs/1702.08608>

- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). **Explainable AI for Trees: From Local Explanations to Global Understanding** (No. arXiv:1905.04610). arXiv. <http://arxiv.org/abs/1905.04610>

Modul:

Gestaltungsziele für menschzentrierte KI

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

01

Einleitung

[Kursübersicht](#) > [Gestaltungsziele für menschzentrierte KI](#)

Künstliche Intelligenz beeinflusst zunehmend, wie Menschen mit automatisierten Systemen interagieren - ob in der Medizin, im Personalwesen, in der Verwaltung oder in Alltagsanwendungen. Doch je autonomer und komplexer KI-Systeme werden, desto wichtiger wird ihre **Gestaltung aus Sicht der Nutzer:innen**. In diesem Modul wollen wir ihnen näher bringen, welche Rolle die verschiedenen Eigenschaften von KI-Systemen spielen und worauf sie bei der Gestaltung ihres Systems achten müssen.

Warum sind UX-Eigenschaften wichtig?

Das folgende Modul bietet einen systematischen Einstieg in zentrale **UX-bezogene Eigenschaften von KI-Systemen**, die darüber entscheiden:

- wie gut Menschen die **Funktion und Grenzen** der KI verstehen,
- ob sie der KI **angemessen vertrauen** (kein blindes oder skeptisches nutzen),
- ob sie im **kritischen Moment handlungsfähig** bleiben,
- und ob sie langfristig die **Verantwortung behalten**.

Diese Eigenschaften sind damit zentrale Voraussetzungen für **Human-Centered AI** und ein entscheidener Faktor für die **gesellschaftliche Akzeptanz und Sicherheit** von KI-Systemen.

Aufbau der Lernmodule

Die behandelten UX-bezogenen Eigenschaften von KI-Systemen sind:

1

Vertrauenswürdigkeit

Wann empfinden Menschen eine KI als glaubwürdig und verlässlich? Dieses Kapitel betrachtet Dimensionen wie Sicherheit und Erklärbarkeit und zeigt, wie UX-Design hilft, Nutzervertrauen richtig zu kalibrieren und Fehlentscheidungen zu vermeiden.

2

Transparenz

Wie kommt eine KI zu ihrem Ergebnis? Dieses Kapitel zeigt, wie Sie technische Prozesse in verständliche Erklärungen übersetzen, damit Nutzer Entscheidungen der KI wirklich verstehen und bewerten können.

3

Erklärbare KI (XAI)

Was macht eine gute Erklärung aus - und für wen? Dieses Kapitel beleuchtet Chancen und Risiken von XAI und zeigt Methoden, um Nutzerkompetenz zu stärken, statt blindes Vertrauen in KI-Systeme zu erzeugen.

4

Kontrollierbarkeit

Wie verhindern wir blindes Vertrauen in KI? Dieses Kapitel zeigt, wie Sie Interfaces gestalten, die Herausforderung die es gibt und wie Sie Nutzern die nötige Kontrolle geben, um Fehler rechtzeitig zu korrigieren.

5

Mentale Modelle

Wie denken Menschen über Systeme - und warum ist das entscheidend? Mentale Modelle sind innere Repräsentationen, die unser Handeln steuern. Erfahren Sie, warum diese Modelle für eine sichere Vorhersage von Systemverhalten entscheidend sind.

Jede Lektion führt in eine zentrale Eigenschaft ein, liefert **praxisnahe Beispiele**, benennt **psychologische und technologische Hintergründe** und bietet **konkrete Empfehlungen für Gestaltung und Umsetzung**.

Das Ziel: Ein fundiertes Verständnis dafür, wie KI-Systeme gestaltet sein müssen, damit sie im Sinne der Menschen funktionieren.

Vertrauenswürdigkeit

Kursübersicht > Gestaltungsziele für menschzentrierte KI

Dieses Kapitel behandelt, warum Vertrauenswürdigkeit für den erfolgreichen und verantwortungsvollen Einsatz von KI-Systemen entscheidend ist, erläutert den Unterschied zwischen Vertrauen und Vertrauenswürdigkeit und zeigt, wie beides durch Gestaltung, Technik und Evaluation gefördert werden kann.

Im folgenden Video wird grundlegend erläutert, was Vertrauenswürdigkeit in KI-Systemen bedeutet und warum sie eine zentrale Voraussetzung für ihre Akzeptanz und verantwortungsvolle Nutzung ist.



<https://youtu.be/aZZJB2xuY88>

1. Einführung: Warum ist Trustworthy AI ein zentrales Thema?

Vertrauenswürdigkeit ist eine Schlüsseldimension für die erfolgreiche Einführung und nachhaltige Nutzung von KI-Systemen. Während technische Leistungsfähigkeit die Funktionsweise bestimmt, entscheidet unter anderem die Wahrnehmung der Vertrauenswürdigkeit darüber, ob Menschen ein System akzeptieren, verantwortungsvoll nutzen und langfristig beibehalten. Es ist daher wichtig, dass Systeme über

Mechanismen oder Merkmale verfügen, die Menschen erkennen lassen, wie vertrauenswürdig sie generell oder in bestimmten Entscheidungen sind. Vertrauenswürdigkeit wird deswegen auch von vielen gesellschaftlichen Initiativen und **Expertenkommissionen** eingefordert.

Besonders in sensiblen Bereichen - etwa in der Medizin, im Finanzwesen oder bei öffentlicher Verwaltung - kann fehlende Vertrauenswürdigkeit gravierende Folgen haben:

- **Gesellschaftlich:** Verlust von Legitimität, Widerstand gegen neue Technologien
- **Individuell:** Fehlentscheidungen durch unberechtigtes Misstrauen in KI-Systeme
- **Wirtschaftlich:** Reputationsschäden, regulatorische Sanktionen, Marktverluste

Internationale Organisationen wie die EU, OECD und IEEE definieren *Vertrauenswürdige KI* als Systeme, die nicht nur funktional, sondern auch **rechtlich, ethisch und technisch** korrekt arbeiten. Der EU AI Act nennt explizit Anforderungen wie Transparenz, Fairness, Sicherheit und menschliche Aufsicht als Kernkriterien. Diese Prinzipien spielen für gemeinwohlorientierte Organisationen eine zentrale Rolle - auch über KI hinaus. Die Risiken beim Einsatz von nicht vertrauenswürdiger KI - oder KI, die zumindest so wirkt - sind daher erheblich.

Vertrauenswürdigkeit ist zunächst eine technische Eigenschaft, die von der Aufgabe der KI und den Zielen des Nutzers abhängt. In ihrer Komplexität ist sie aber ein **interdisziplinäres Gestaltungsziel**, das technologische, regulatorische und UX-bezogene Aspekte vereint und nicht nur die Optimierung der Leistung eines Systems beinhaltet,

sondern auch Menschen die Möglichkeit geben soll, das System einzuschätzen.

Aufgrund der Komplexität des Begriffs Vertrauenswürdigkeit ist eine Definition nicht einfach. Im nächsten Abschnitt widmen wir uns deshalb der Frage, **warum es schwierig ist, Vertrauenswürdigkeit klar zu definieren**, und nähern uns dadurch einer Definition an.

2. Warum ist Vertrauenswürdigkeit schwer zu definieren?

Obwohl sie *objektiv* wirken soll, ist Vertrauenswürdigkeit schwierig allgemein und einheitlich zu definieren, denn:

- Sie besteht aus mehreren Dimensionen (z.B. Transparenz, Fairness, Robustheit).
- Ihre Bewertung ist kontextabhängig (Was im E-Commerce als vertrauenswürdig gilt, reicht im Gesundheitswesen vielleicht nicht aus - der Fachausdruck ist „individueller Standard“).
- Sie wird oft mit Vertrauen verwechselt oder vermischt.

Die Begriffe *Vertrauen* und *Vertrauenswürdigkeit* sind nicht identisch. Gerade aus psychologischer Perspektive lohnt sich die Unterscheidung. Also, wo genau liegen die Unterschiede?

3. Vertrauen vs. Vertrauenswürdigkeit

Merksatz: *Vertrauen ist eine Einstellung, die Menschen haben.*

Vertrauenswürdigkeit ist eine Eigenschaft, die ein System (in einem Kontext) hat.

Der Unterschied zwischen *Vertrauen* und *Vertrauenswürdigkeit* ist zentral für die Gestaltung und Bewertung von KI-Systemen:

1. Vertrauen ist eine **subjektive Haltung** bzw. Einstellung eines Individuums oder einer Gruppe gegenüber einer Entität (hier: der KI). Es basiert auf Wahrnehmung, Erfahrung, Intuition und oft auch auf psychologischen und kulturellen Faktoren. Vertrauen kann entstehen, selbst wenn ein System objektiv unsicher ist - oder ausbleiben, obwohl das System technisch und ethisch einwandfrei funktioniert.

2. Vertrauenswürdigkeit ist eine **objektive, überprüfbare Eigenschaft des Systems.**

Sie hängt von Kriterien wie Zuverlässigkeit, Fairness, Sicherheit, Transparenz und Erklärbarkeit ab. Ein vertrauenswürdiges System erfüllt dokumentierte Standards und kann seine Leistungsfähigkeit und Unvoreingenommenheit nachweisen.

Eine Verwechslung ist leicht möglich: Vertrauenswürdigkeit kann nämlich in einer Anwendung mit wenig Risiko und Anspruch an Korrektheit schneller gegeben sein, als in einem Kontext, in dem Fehler sehr gefährlich sind. Dadurch wirkt Vertrauenswürdigkeit aufgrund ihrer Kontextabhängigkeit nicht überprüfbar und objektiv - so wie Vertrauen.

3a) Wie entsteht Vertrauen in KI-Systeme?

Vertrauen entsteht **nicht automatisch** durch technische Qualität. Es ist ein psychologischer und sozialer Prozess. Ein hilfreiches Modell zur Beschreibung dieses Prozesses stammt aus der Forschung zur Mensch-Computer-Interaktion. **Madsen und Gregor (2000)** unterscheiden darin zwei zentrale Dimensionen von Vertrauen in Computersysteme:

Kognitives Vertrauen

Beruht auf der rationalen Einschätzung der Systemleistung. Es entsteht, wenn Nutzer:innen das System als kompetent, vorhersehbar und zuverlässig wahrnehmen.

Fördernde Faktoren:

- technische Kompetenz und Genauigkeit
- konsistente, nachvollziehbare Entscheidungen
- transparente Abläufe
- Stabilität und Verlässlichkeit im Betrieb

Affektives Vertrauen

Beruht auf emotionaler Resonanz und sozialer Wahrnehmung. Es entsteht, wenn Nutzer:innen das Gefühl haben, fair behandelt zu werden oder dass das System ihre Interessen unterstützt.

- menschlich wirkendes, empathisches Design - aber Achtung, es sollte kein *uncanny valley* entstehen
- freundliche, respektvolle Sprache und soziale Signale
- ethisches Verhalten (z.B. keine Manipulation, kein übertriebener Druck)

UX-Design muss beide Dimensionen - kognitiv und affektiv - mitdenken, um angemessenes Vertrauen in KI-Systeme zu ermöglichen.

3b) Warum reicht Vertrauen allein nicht aus?

Ein entscheidender Punkt: Nur weil Menschen einem System vertrauen, ist es noch lange nicht vertrauenswürdig. Und umgekehrt vertrauen Menschen einem System nicht direkt, nur weil es Vertrauenswürdig ist.

- 1. Risiko:** Menschen vertrauen einem **nicht vertrauenswürdigen** System
 - Gefahr von Fehlentscheidungen.

Beispiel: Nutzende vertrauen einem nicht für medizinische Beratung ausgelegten System wie ChatGPT bei Fragen zu komplexen Wechselwirkungen von Medikamenten. Dies nennt man Übervertrauen.

- 2. Risiko:** Menschen misstrauen einem **vertrauenswürdigen** System
 - Gefahr von Ineffizienz, Ablehnung, Algorithm Aversion.

Beispiel: Es wird lieber manuell ein komplexer Datensatz aufgearbeitet, als sich auf ein automatisiertes System zu verlassen, das für diese Aufgabe geschaffen worden ist. Dies nennt man Untervertrauen.

Deshalb ist das Ziel von UX-Design und KI-Entwicklung:
Vertrauenswürdigkeit sicherstellen (systemseitig) und **Vertrauen kalibrieren (nutzerseitig)**.

4. Dimensionen vertrauenswürdiger KI- Systeme

Für ein System, das als vertrauenswürdig gelten soll, werden in der Regel folgende Eigenschaften gefordert:

a) Technische Robustheit und Sicherheit

Das System soll unter normalen und außergewöhnlichen Bedingungen zuverlässig arbeiten. Zu relevanten Aspekten zählen z. B. Fehlertoleranz, Resilienz gegen Angriffe (Cybersecurity), Fail-Safe-Mechanismen, kontinuierliche Überwachung.

UX-Bezug: Nutzer:innen müssen über Systemstatus, Ausfälle oder Sicherheitsereignisse klar informiert werden.

b) Transparenz und Erklärbarkeit

Entscheidungen und Prozesse sollen nachvollziehbar und überprüfbar sein. Dazu zählt u. a. die Offenlegung der Funktionsweise (z. B. Modellarchitektur, Trainingsdatenquellen), Erklärungen einzelner Entscheidungen, Angabe von Unsicherheiten.

UX-Bezug: Erklärungen müssen in für die Zielgruppe verständlicher Form präsentiert werden (Text, Visualisierung, interaktive Elemente).

c) Fairness

KI soll Personen oder Gruppen nicht benachteiligen oder privilegieren, es sei denn, dies ist explizit gerechtfertigt (z. B. positive Diskriminierung). Dazu gehört u. a. Bias-Erkennung, faire Datenauswahl, Überprüfung von Outputs auf diskriminierende Muster.

UX-Bezug: Betroffene müssen bei Ergebnissen erkennen und nachvollziehen können, ob diese aufgrund verzerrter Daten zustande gekommen sind.

d) Datenschutz und Daten-Governance

Schutz personenbezogener Daten und verantwortungsvoller Umgang mit sensiblen Informationen. Dazu zählt u. a. Privacy by Design, Minimierung erhobener Daten, klare Einwilligungsprozesse, Datenanonymisierung.

UX-Bezug: Nutzer:innen müssen leicht nachvollziehen und steuern können, welche Daten genutzt werden.

e) Rechenschaftspflicht & Verantwortung

Es muss klar sein, wer für das Verhalten des Systems verantwortlich ist, und es muss möglich sein, Entscheidungen im Nachhinein zu überprüfen. Dazu gehört z. B. Dokumentation, Audit-Trails, klare Verantwortlichkeitszuordnung, Haftungsregelungen.

UX-Bezug: Nutzer:innen müssen wissen, an wen sie sich im Falle von Problemen wenden können.

f) Human Agency & Oversight

Menschen behalten die Kontrolle über kritische Entscheidungen. Dazu zählen z. B. Mechanismen wie Human-in-the-Loop, Abschaltmöglichkeiten, Entscheidungsunterstützung statt -ersetzung.

UX-Bezug: Schnittstellen müssen Eingriffe intuitiv ermöglichen, ohne dass Nutzer:innen durch komplexe Prozesse abgeschreckt werden.

Diese Dimensionen bilden das Fundament der objektiven Vertrauenswürdigkeit. UX-Design hat die Aufgabe, diese Eigenschaften **erlebbar** zu machen, sodass sie nicht nur technisch vorhanden sind, sondern auch subjektiv wahrgenommen werden.

5. Was folgt daraus für die Gestaltung von KI?

Empfehlung für die Praxis:

1

Entwickeln Sie **technisch vertrauenswürdige Systeme**, die fair, robust und nachvollziehbar sind.

2

Gestalten Sie **erklärende Interfaces**, die Nutzer:innen wirklich verstehen können.

3

Testen Sie mit echten Nutzer:innen: **Verstehen ihre Nutzer:innen die Entscheidungen des Systems?**

4

Kommunizieren Sie ehrlich: **Keine Überversprechen von KI-Fähigkeiten!**

Aber wie lassen sich Vertrauenswürdigkeit und Vertrauen in Bezug auf ein KI-System eigentlich messen?

6. Messung von Vertrauen und Vertrauenswürdigkeit

Die Evaluation muss zwischen **subjektivem Vertrauen** und **objektiver Vertrauenswürdigkeit** unterscheiden. Diese beiden Maße können auseinanderfallen und sollten separat erhoben werden.

Messung von Vertrauen (subjektiv)

- **Umfragen & Fragebögen:** z. B. Trust in Automation Scale, NASA-TLX (für mentale Belastung)
- **Verhaltensindikatoren:** Bspw. Häufigkeit, mit der Nutzer:innen Empfehlungen der KI folgen oder sie ablehnen
- **Langzeitbeobachtung:** Veränderungen des Vertrauens über wiederholte Nutzung

Messung von Vertrauenswürdigkeit (objektiv)

- **Technische Metriken:** Genauigkeit, Fehlerraten, Fairness-Indikatoren, Robustheitstests
- **Audit & Compliance-Prüfungen:** Abgleich mit regulatorischen Standards (z. B. EU AI Act, ISO-Normen)
- **Erklärbarkeits-Checks:** Verständlichkeit und Korrektheit der bereitgestellten Erklärungen

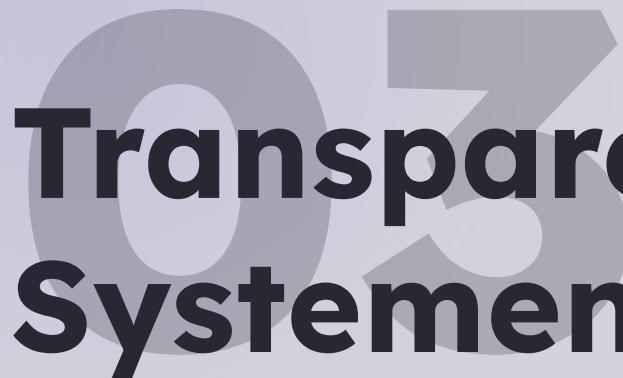
Kombination von Messmethoden

Gemeinsame Auswertung, um *Trust Calibration* zu prüfen - also ob subjektives Vertrauen mit objektiver Vertrauenswürdigkeit übereinstimmt.

7. Fazit: Vertrauen gestalten, Vertrauenswürdigkeit sichern

Vertrauenswürdigkeit ist kein Marketing-Schlagwort, sondern eine **gestalterische Verantwortung**. Sie verlangt technisches Know-how, psychologisches Verständnis und ethische Klarheit.

Die Frage ist nicht: Wie überzeugen wir Menschen von KI?
Sondern: Wie gestalten wir KI, die überzeugt?

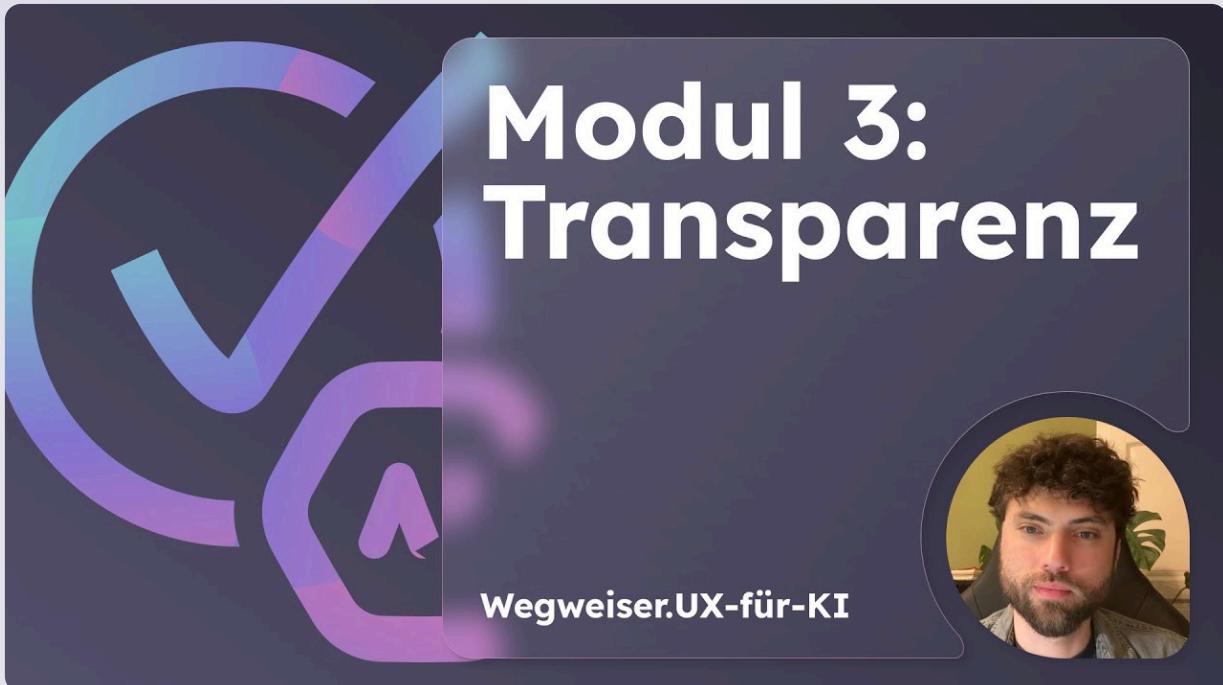


Transparenz in KI-Systemen

Kursübersicht > Gestaltungsziele für menschzentrierte KI

Dieses Kapitel beleuchtet die Bedeutung von Transparenz als zentrale Voraussetzung vertrauenswürdiger KI und zeigt, warum sie nicht nur rechtlich gefordert, sondern auch essenziell für Verständnis, Verantwortung und Akzeptanz ist.

Im folgenden Video wird Transparenz von KI-Systemen anhand eines Beispiels erklärt und darauf eingegangen wie man Transparenz erreichen kann, sowie dessen Aspekte im Bezug zur UX.



<https://youtu.be/yRCWVsM9SAY>

1. Einleitung: Warum ist Transparenz wichtig?

Transparenz ist ein zentrales Prinzip im Kontext vertrauenswürdiger KI. Sie wird oft als Voraussetzung dafür genannt, dass ein System als **vertrauenswürdig** wahrgenommen werden kann. Aber: **Transparenz ist nicht das Gleiche wie Vertrauenswürdigkeit**. Sie ist vielmehr eine **notwendige Bedingung** für korrekte Vertrauenswürdigkeit und viele andere Konstrukte wie Erklärbarkeit, Kontrollierbarkeit und letztlich damit die Nützlichkeit von KI-Systemen.

Besonders im Kontext wachsender regulatorischer Vorgaben - etwa dem **EU AI Act**, der für KI-Systeme mit annehmbaren Risiko explizit Transparenzpflichten vorsieht - ist Transparenz nicht nur eine Frage des Vertrauens, sondern eine rechtliche Notwendigkeit.

2. Definition: Was bedeutet Transparenz in der KI?

Transparenz beschreibt die Fähigkeit eines Systems, für Nutzer:innen **einsichtig, nachvollziehbar und interpretierbar** zu sein. Nutzer:innen können einsehen:

- **Wie** kommt die KI zu ihrer Entscheidung?
- **Welche Daten** wurden verwendet?
- **Welche Annahmen** und Verfahren liegen dem Modell zugrunde?
- **Welche Grenzen**, Unsicherheiten, Verzerrungen bestehen?
- **Welche Ziele hat das System?**

Transparenz kann dabei auf **verschiedenen Ebenen** stattfinden:

- **Technisch** (Code, Algorithmen, Trainingsdaten)
- **Funktional** (Input-Output-Zusammenhang)
- **Erklärend** (für Nutzende nachvollziehbar)
- **Organisatorisch** (Verantwortlichkeiten, Dokumentation)

3. Beispiel: Denkmal-Entscheidung per Bildanalyse

Ein KI-System soll anhand eines Fotos entscheiden, ob ein Gebäude denkmalgeschützt ist. Das System erklärt:

„Wir haben 500 Bilder von denkmalgeschützten Gebäuden und 500.000 Bilder von nicht denkmalgeschützten Gebäuden verwendet.“

Obwohl diese Information **technisch korrekt** ist, zeigt sie nur teilweise Transparenz:

- Ein **extremes Datenungleichgewicht** kann zu systematischer Verzerrung (Bias) führen. Es ist unklar, ob dies überprüft wurde.
- Die Information liefert **keinen Einblick in die eigentliche Entscheidungslogik** des Systems.
- Nutzer:innen erhalten zwar Informationen, aber nicht das, was für eine **vertrauenswürdige Bewertung** der Entscheidung wichtig wäre.

Merksatz: *Transparenz ist nur dann hilfreich, wenn sie die für Nutzer:innen relevanten Aspekte sichtbar macht.*

4. Warum ist Transparenz nicht trivial?

Transparenz ist komplex - sowohl in der technischen Umsetzung als auch in der UX-Vermittlung. Dabei gibt es einige häufige Missverständnisse:

- „Transparenz = Offenlegen von Code“ - für die meisten Nutzer:innen, die nicht technisch versiert sind, ist eine solche Information **nicht hilfreich**
- „Mehr Transparenz ist immer besser“ - kann aber auch zu **Verwirrung oder Misstrauen führen**

Herausforderungen:

- Unterschiedliche Zielgruppen benötigen **unterschiedliche Erklärungen** (z.B. Laien vs. Expert:innen)
- Komplexe Modelle (z. B. Deep Learning) lassen sich nicht immer erklären
- Zielkonflikte: Transparenz vs. Datenschutz, Sicherheit, geistiges Eigentum

Was passiert ohne zielgerichtete Transparenz?

- Nutzer:innen erhalten Daten - aber nicht **nicht das, was sie brauchen**, um Entscheidungen zu bewerten
- Transparenz verkommt zur **Schein-Transparenz**, wenn Relevanz fehlt
- Besonders kritisch bei Verzerrungen in Trainingsdaten oder Black-Box-Verfahren

5. Wie lässt sich Transparenz herstellen?

Transparenz ist kein einmaliger Zustand, sondern ein Designprozess.

Sie lässt sich auf mehreren Ebenen gestalten:

- **Daten & Entwicklung dokumentieren:** Welche Daten wurden verwendet? Wie wurden sie bereinigt oder gefiltert?
- **Beteiligte offenlegen:** Wer hat das System entwickelt? Welche Interessen könnten beeinflusst haben? Was für Stakeholder gab es?
- **Ressourcen transparent machen:** Wie viel Energie, Rechenleistung oder Zeit wurde aufgewendet?
- **Ergebnisse kommunizieren:** Wie zuverlässig sind die Vorhersagen? Wie wurden sie validiert?

UX-bezogene Transparenz: Was brauchen Nutzer:innen?

Aus UX-Sicht geht es nicht nur um Offenlegung, sondern um **verstehbare Darstellung**. Ziel ist es, Nutzer:innen die Möglichkeit zu geben, **die Entscheidungen des Systems sinnvoll einzuordnen**.

Wichtige UX-Fragen zur Transparenz

- Welche Daten nutzt die KI - und warum?
- Wie wurde das Modell trainiert?
- Wie kommt das System zu seinem Ergebnis?
- Wie zuverlässig ist dieses Ergebnis?
- Welche Grenzen, Risiken oder Unsicherheiten bestehen?

Drei Arten von Transparenz

1. Prozess-Transparenz

Offenlegung der Entstehung und Funktionsweise eines KI-Systems.

- **Beispieleweise:** Herkunft und Qualität der Trainingsdaten, Beschreibung der Modellarchitektur und verwendeten Algorithmen, inklusive Trainingsverhalten, sowie Zieldefinition und Systemgrenze.
- **Zweck:** Hilft Nutzer:innen und Prüfern, konkrete Ergebnisse zu verstehen, zu hinterfragen und ggf. zu korrigieren

UX-Bezug: Prozessinformationen müssen in einer Form verfügbar sein, die sowohl für Fachleute als auch für betroffene Nutzer:innen zugänglich ist - z.B. über interaktive Dokumentationen oder „About this AI“-Sktionen.

2. Entscheidungs-Transparenz

Nachvollziehbarkeit einzelner Entscheidungen oder Outputs der KI

- **Beispielsweise:** Begründung, warum eine bestimmte Entscheidung getroffen wurde, Darstellung der wichtigsten Einflussfaktoren, Angabe von Unsicherheiten oder Wahrscheinlichkeiten
- **Zweck:** Hilft Nutzer:innen und Prüfern, konkrete Ergebnisse zu verstehen, zu hinterfragen und ggf. zu korrigieren

UX-Bezug: Erklärungen müssen kontextbezogen und handlungsrelevant sein, z. B. durch visuelle Hervorhebung relevanter Datenpunkte oder Szenario-abhängige Erklärtexete.

3. Governance-Transparenz

Offenlegung der organisatorischen und regulatorischen Rahmenbedingungen, unter denen ein KI-System betrieben wird.

- **Beispielsweise:** Zuständigkeiten und Verantwortlichkeiten, Eingesetzte Audit- und Überwachungsprozesse, Einhaltung von Standards und Zertifizierungen
- **Zweck:** Ermöglicht es Stakeholdern, die Verantwortungsstruktur zu verstehen und im Problemfall geeignete Ansprechpersonen zu finden

UX-Bezug: Governance-Informationen sollten für Endnutzer:innen einfach auffindbar sein, z. B. über leicht zugängliche Hilfeseiten, Zertifikatsanzeigen oder Compliance-Labels im Interface.

Zusammenhang der Dimensionen:

1

Prozess-Transparenz → zeigt **wie** das System gebaut ist

2

Entscheidungs-Transparenz → erklärt **warum** das System etwas tut

3

Governance-Transparenz → offenbart **wer** dafür verantwortlich ist

Alle drei Dimensionen zusammen ermöglichen nicht nur eine **objektive Nachvollziehbarkeit**, sondern auch eine **subjektive Vertrauensbildung** - vorausgesetzt, sie werden verständlich aufbereitet.

6. Praktische Tipps zur Gestaltung transparenter KI

1

Kenntnis der Zielgruppe: Was wollen Nutzer:innen wissen? Was können sie verstehen?

2

Relevanz statt Überfrachtung: Nur die Informationen geben, die für die Entscheidung oder Nutzung wichtig sind.

3

Visuelle Unterstützung: Erklärungen durch Diagramme, Heatmaps, Gegenbeispiele etc.

4

Transparenz modularisieren: Für verschiedene Ebenen (Daten, Modell, Entscheidung) unterschiedliche Erklärungstiefen anbieten.

5

Feedback einholen: Verstehen die Nutzer:innen wirklich, was erklärt wurde?

7. Fazit: Transparenz als Brücke zum Vertrauen und Grundlage für weitere UX-bezogene Eigenschaften

Transparenz ist kein Selbstzweck - sie ist ein **ethisches und nutzerzentriertes Gestaltungsprinzip** das:

- Vertrauen aufbaut
- Verantwortung ermöglicht
- Erklärbarkeit, Kontrollierbarkeit und Nützlichkeit unterstützt
- Missverständnisse und Fehlverhalten verhindert

Gute Transparenz ist immer adressatengerecht, relevant und handlungsunterstützend. Sie ist die Brücke zwischen komplexer Technik und verständlicher, verantwortungsvoller Nutzung.

Erklärbare KI (XAI)

[Kursübersicht](#) > [Gestaltungsziele für menschzentrierte KI](#)

Dieses Kapitel führt in das Konzept der Erklärbaren Künstlichen Intelligenz (XAI) ein und zeigt, warum Erklärbarkeit entscheidend ist, um Vertrauen, Verantwortung und Verständnis im Umgang mit KI-Systemen zu fördern.

Im folgenden Video wird anhand eines Beispiels grundlegend erklärt was Erklärbare KI ist.



<https://youtu.be/OyxclsV4ysE>

1. Einleitung: Was ist XAI und warum ist sie wichtig?

Erklärbare Künstliche Intelligenz (engl. Explainable AI, XAI) bezieht sich auf technische Lösungen und Strategien, die es Menschen ermöglichen, die Entscheidungen und Funktionsweise von KI-Systemen nachzuvollziehen. Wieso benötigen wir das überhaupt? In verschiedenen Fällen reicht es nicht aus, dass eine KI zuverlässig funktioniert - sie muss auch erklären können, warum sie zu einer bestimmten Entscheidung gelangt ist.

Diese Erklärbarkeit ist zentral für:

- den **Aufbau von Vertrauen** in automatisierten Systemen,
- die **Verantwortungszuschreibung** bei fehlerhaften Entscheidungen,
- die **Nutzungskompetenz** bei Endanwender:innen,
- die **Erfüllung gesetzlicher Anforderungen**, z.B. durch den EU Act,
- sowie die Entwicklung **mentaler Modelle**, die Nutzer:innen helfen, ein System korrekt zu interpretieren und angemessen zu nutzen.

Miller (2019) betont dabei, dass Erklärbarkeit keine rein technische Transparenz ist. Vielmehr geht es darum, dass Menschen eine Entscheidung nachvollziehen können - auf eine Art, die für sie verständlich und bedeutsam ist.

2. Sind KI-Systeme immer erklärbar?

Nicht jedes KI-System lässt sich einfach erklären. Während regelbasierte Systeme oder klassische statistische Modelle oft relativ durchschaubar sind, stoßen wir bei modernen KI-Ansätzen schnell an Grenzen der Verständlichkeit. Besonders **Deep-Learning-Modelle**, die heute in vielen Anwendungen wie Bilderkennung, Sprachverarbeitung oder Empfehlungssystemen eingesetzt werden, gelten häufig als „**Black Boxes**“. Deep Learning basiert auf **künstlichen neuronalen Netzen**, die aus vielen Schichten („Layers“) miteinander verbundener künstlicher Neuronen bestehen und so hochkomplexe Muster und Zusammenhänge in großen Datenmengen automatisch erkennen und verarbeiten können.

Was bedeutet „Black Box“?

Eine „Black Box“ beschreibt ein System, dessen **innere Entscheidungsprozesse für Menschen nicht direkt nachvollziehbar** sind. Zwar können wir die Eingaben und Ausgaben eines Modells sehen, aber die Vielzahl an internen Berechnungen bleibt verborgen.

Warum entsteht die Black-Box-Problematik?

- **Komplexität der Modelle:** Deep-Learning-Netzwerke bestehen oft aus Millionen oder sogar Milliarden Parametern, die in vielen Schichten (Layers) organisiert sind.
- **Nichtlineare Zusammenhänge:** Diese Netzwerke lernen hochkomplexe Muster, die sich nicht einfach in Regeln übersetzen lassen.
- **Automatisches Feature-Learning:** Anders als bei klassischen Modellen werden relevante Merkmale (Features) nicht von Menschen vorgegeben, sondern automatisch gelernt - was Transparenz erschwert.
- **Optimierungsverfahren:** Trainingsprozesse wie Gradient Descent optimieren die Parameter, ohne dass für Menschen intuitive Zusammenhänge sichtbar sind.

Konsequenz

Die Black-Box-Natur moderner KI-Modelle macht es schwierig, **Erklärbarkeit** und **Nachvollziehbarkeit** zu gewährleisten. Das bedeutet jedoch nicht, dass Erklärbarkeit unmöglich ist: Mit Methoden wie Feature-Attribution, Modellvereinfachungen oder Interpretable-by-Design-Ansätzen gibt es Werkzeuge, die Licht ins Dunkel bringen.

3. Warum ist Erklärbarkeit komplex?

Eine Erklärung im Kontext von XAI ist ein kommunikatives Mittel, um auf Fragen wie „Warum wurde diese Entscheidung getroffen?“, „Was hätte passieren müssen, damit es anders kommt?“ oder „Was war besonders einflussreich?“ eine verständliche Antwort zu geben.

Allerdings ist das Konzept „Erklärung“ schwer zu fassen, denn:

- **Kontextabhängigkeit:** Je nach Anwendung (Medizin, Kreditvergabe, Bildklassifikation) ändern sich die Anforderungen an die Erklärung.
- **Zielgruppenunterschiede:** Fachleute, Endnutzende und Aufsichtsbehörden benötigen unterschiedliche Formate und Tiefen.
- **Technisch korrekt ≠ kognitiv hilfreich:** Eine präzise technische Begründung hilft nur, wenn sie verstanden wird.

Ein Beispiel: Die Aussage „Die Entscheidung basiert auf der Position des Entscheidungsraums in Feature X“ ist für Laien nicht hilfreich. Besser wäre: „Ihr monatliches Einkommen liegt unter 3.200€, was zur Ablehnung beigetragen hat.“

4. Arten von Erklärungen in XAI

Erklärungen können auf unterschiedliche Weisen strukturiert sein. Man unterscheidet insbesondere **Lokale und Globale Erklärungen**:

Lokale Erklärungen

Diese beziehen sich auf eine **konkrete Entscheidung** eines KI-Systems. Sie beantworten die Frage: „Warum genau wurde in diesem Fall X und nicht Y entschieden?“

- Zeigen den Einfluss einzelner Eingabeparameter
- Typische Methoden: SHAP, LIME, Counterfactuals

Globale Erklärungen

Sie beschreiben die **allgemeine Funktionsweise** des Modells über viele Entscheidungen hinweg:

- Sie geben einen Überblick über Entscheidungslogik des Modells und erläutern,
- welchen Einfluss verschiedene Variablen haben und wie sie zusammenhängen

Beispiel für ein Kreditbewertungsmodell

Ein Kreditbewertungsmodell (Scoring-Modell) wird global analysiert, um zu verstehen, welche Faktoren insgesamt am stärksten die Kreditwürdigkeit beeinflussen.

Die globale Erklärung zeigt z. B.:

- Einkommen hat hohen positiven Einfluss auf den Score.
- Hohe Kreditkartenauslastung wirkt sich negativ aus.
- Alter spielt nur eine geringe Rolle.

Wechselwirkungen: Hohe Auslastung und niedriges Einkommen verstärken den negativen Effekt.

So wird deutlich, welche Muster das Modell generell gelernt hat, unabhängig von einer einzelnen Kundenentscheidung.

Weitere Einteilungen (Speith, 2020)

Post-hoc vs. intrinsisch: Erklärung wird entweder nachträglich erzeugt oder ergibt sich aus der Modellstruktur selbst (z.B. Entscheidungsbaum).

Modellbasierte (intrinsische) Erklärbarkeit

- **Entscheidungsbäume** - Entscheidungen folgen klaren Regeln
- **Lineare Modelle** - Einfluss jedes Faktors ist direkt absehbar
- **Regel- oder logikbasierte Systeme** - nachvollziehbare IF-THEN-Strukturen

Post-hoc-Erklärungen

Hier wird das Verhalten eines komplexen, intransparenten Modells nachträglich analysiert. Häufige Ansätze sind:

- **Feature-Attribution:** Wie wichtig war ein bestimmtes Eingabefeature für diese Entscheidung?
 - **SHAP (SHapley Additive exPlanations)**
 - **LIME (Local Interpretable Model-Agnostic Explanations)**
- **Kontrastive Erklärung:** Warum wurde A statt B vorhergesagt?
- **Gegenfaktische Erklärung:** Was müsste sich an den Eingabedaten ändern, damit B statt A passiert?
- **Symbolisch vs. visuell:** Textlich formuliert vs. visuelle Hilfsmittel wie Diagramme, Heatmaps, Salience Maps

5. Wirkung von Erklärungen - Chancen und Risiken

Erklärungen können gut gestaltet sein, wobei es deutliche Grenzen gibt und sie auch so gestaltet sein können, dass es problematisch ist.

Gut gestaltete Erklärungen

- **Verständlichkeit:** Klar, nachvollziehbar, ohne Fachjargon
- **Relevanz:** Fokussiert auf das, was Nutzer:innen wirklich interessiert
- **Treffsicherheit:** Erfasst die zentrale Logik der Entscheidungen
- **Vertrauensbildung:** Fördert angemessenes Vertrauen (weder blind noch misstrauisch)
- **Lernförderlich:** Hilft, ein mentales Modell aufzubauen

Grenzen

- **Komplexität des Modells:** Hochdimensionale Netze haben keine klaren "Entscheidungswege"
- **Datenabhängigkeit:** Erklärungen sind nur so gut wie die Daten, die verwendet wurden
- **Missverständnisse:** Nutzer:innen interpretieren Erklärungen anders als intendiert
- **Manipulation:** Erklärungen können auch genutzt werden, um Vertrauen zu erzwingen

Problematisch

- **Falsche oder ungenaue Erklärungen** können zu fehlerhaften Verhalten führen
- **Übermäßige Vereinfachungen** können relevante Aspekte verschleiern
- **Erklärungen können manipulativ wirken**, wenn sie Vertrauen erzeugen sollen, wo Misstrauen angemessen wäre

Beispiel aus der Forschung (Kühl et al., 2024)

In einem Experiment zu Altersschätzungen zeigte sich:
Teilnehmende vertrauten einem System mehr, wenn es eine
plausibile Erklärung (egal ob richtig oder falsch) lieferte - selbst
wenn die Entscheidung objektiv falsch war. Das bedeutet: **Eine gut
präsentierte, aber falsche Erklärung kann gefährlicher sein als
keine Erklärung.**

6. Gestaltungshinweise und praktische Tipps

Damit Erklärbarkeit in der Praxis gelingt, sollten folgende Grundsätze beachtet werden:

Planung und Einbindung

- **Frühzeitig mitdenken:** XAI sollte integraler Bestandteil der Entwicklung sein
- **Zielgruppe definieren:** Welche Fragen stellen die Nutzer:innen? Was wollen sie wirklich wissen?
- **Mit Nutzenden gemeinsam definieren,** was erklärt werden soll (z. B. mit dem Question-Driven Design nach Liao et al., 2021)
- **Kontextabhängig gestalten:** Je nach Anwendung und Zielgruppe andere Erklärungen
- **Exploration statt einfache Aussagen:** Nutzer:innen sollen selbst Zusammenhänge entdecken können
- **Auf mentale Modelle achten:** Wie denken die Nutzer:innen über die KI?
- **Vermeiden von Overtrust:** Nicht alles erklären, was das Modell tut, sondern nur das, was sinnvoll und hilfreich ist

Methoden und Darstellung

- **Visualisierung nutzen:** z.B. Feature-Highlights, Balkendiagramme, Overlay-Heatmaps
- **Mehrere Erklärungstypen anbieten:** Für verschiedene Nutzungskontexte
- **Exploration ermöglichen:** Nutzer:innen sollen nicht nur konsumieren, sondern auch interaktiv verstehen können

Evaluation und Feedback

Fragen Sie Ihre Nutzer:innen: Was möchten Sie wissen? Warum ist diese Entscheidung relevant für Sie?

Nutzen Sie einfache Visualisierungen, z.B. Feature-Highlights, Balken, Overlay-Grafiken Testen Sie Ihre Erklärungen mit realen Nutzenden und beobachten Sie, ob deren Verhalten sich verbessert.

- **Iterativ testen** mit echten Nutzenden
- **Verstehen evaluieren**, nicht nur Zufriedenheit
- **Erklärungsnutzung beobachten:** Wird erklärt, aber nicht verstanden?
Wird ignoriert?

7. Fazit: XAI als kontinuierlicher Gestaltungsprozess

Erklärbarkeit ist kein statisches Feature, sondern ein **dynamisches Element der Mensch-KI-Interaktion**. Systeme, Nutzer:innen und Anwendungskontexte entwickeln sich weiter - gute XAI begleitet diesen Wandel.

XAI dient nicht nur der Transparenz, sondern auch der **Wissensvermittlung, Kontrolle und Selbstwirksamkeit**. Eine erklärbare KI ist eine nutzbare und verantwortbare KI.

„Explain unto others in such a way as to help them explain to themselves.“ - Hoffman et al. (2023)

05 Kontrollierbarkeit

[Kursübersicht](#) > [Gestaltungsziele für menschzentrierte KI](#)

Dieses Kapitel behandelt die Kontrollierbarkeit von KI-Systemen aus einer UX-orientierten Perspektive und zeigt, wie Nutzer:innen die Möglichkeit erhalten, das Verhalten von KI gezielt zu verstehen, zu beeinflussen und sicher zu steuern - eine zentrale Voraussetzung für Akzeptanz, Vertrauen und verantwortungsvollen Einsatz.

Im folgenden Video wird grundlegend erläutert, was Kontrollierbarkeit in der Mensch-KI-Interaktion bedeutet und warum sie eine Schlüsselrolle für nutzerzentriertes, sicheres und vertrauenswürdiges KI-Design spielt.



<https://youtu.be/Mu8MafXxgVI>

Grundlagen der Kontrollierbarkeit in KI (UX- orientiert)

1. Einleitung: Kontrollierbarkeit in der Mensch-KI-Interaktion

Kontrollierbarkeit beschreibt die Fähigkeit, das Verhalten eines Systems gezielt zu beeinflussen oder zu begrenzen, sodass es mit den Zielen des

Menschen übereinstimmt. Während dieser Begriff in der klassischen Regelungstechnik vor allem mathematisch definiert ist - etwa als Möglichkeit, ein System aus jedem beliebigen Ausgangszustand in einen gewünschten Endzustand zu überführen - verschiebt sich der Fokus im Kontext moderner KI-Systeme mit direkter Mensch-Maschine-Interaktion deutlich.

In einer UX-orientierten Perspektive geht es weniger um die vollständige mathematische Kontrollierbarkeit des Modells, sondern vielmehr um die *wahrgenommene und erlebbare Kontrollierbarkeit* aus Sicht der Nutzer:innen. Die entscheidenden Fragen lauten:

- **Verstehen** die Nutzer:innen, was die KI tut?
- **Können** sie in den Prozess eingreifen, wenn nötig?
- **Erleben** sie ein angemessenes Maß an Kontrolle, das Vertrauen schafft, ohne die Funktionalität einzuschränken?

Gerade bei KI-Systemen mit zunehmender Autonomie (z. B. generative Sprachmodelle, autonome Fahrzeuge, adaptive Empfehlungssysteme) ist diese Form der Kontrollierbarkeit essenziell für Akzeptanz, Sicherheit und verantwortungsvollen Einsatz. Forschungen in der *Human-Computer Interaction* (HCI) zeigen, dass wahrgenommene Kontrollmöglichkeiten maßgeblich das Vertrauen in automatisierte Systeme beeinflussen. Fehlende Kontrolle - oder auch nur das Gefühl mangelnder Eingriffsmöglichkeiten - führt dagegen häufig zu Ablehnung oder riskantem Verhalten, etwa blindem Vertrauen ohne kritische Prüfung.

Aus UX-Sicht wird Kontrollierbarkeit zu einer *Schnittstellenaufgabe*: Sie hängt nicht nur von der inneren Architektur des KI-Systems ab, sondern stark von der Gestaltung der Interaktionsmöglichkeiten, der Transparenzmechanismen und der Einbettung in den Nutzungskontext.

2. Dimensionen der Kontrollierbarkeit aus UX-Perspektive

Aus Sicht der Mensch-KI-Interaktion lässt sich Kontrollierbarkeit in mehrere zentrale Dimensionen unterteilen. Diese Dimensionen bestimmen, wie gut Nutzer:innen in der Lage sind, die KI zu verstehen, zu beeinflussen und zu überwachen. Sie sind nicht nur technische Eigenschaften, sondern auch Gestaltungsprinzipien für Interfaces und Interaktionsdesign.

a) Transparenz

Transparenz bedeutet, dass das System seine Funktionsweise, Entscheidungslogik und Zielrichtung in einer für den Menschen verständlichen Form offenlegt. In der UX-Praxis umfasst das:

- Klar erkennbare Systemzustände
- Erklärungen zu Entscheidungen (z. B. warum ein bestimmtes Ergebnis vorgeschlagen wird)
- Sichtbare Unsicherheiten oder Grenzen des Systems

Hohe Transparenz erleichtert es, mentale Modelle zu bilden, die Grundlage für effektive Kontrolle sind.

b) Vorhersagbarkeit

Ein KI-System sollte in vergleichbaren Situationen konsistent reagieren. Vorhersagbarkeit verringert die kognitive Belastung, da Nutzer:innen weniger Energie darauf verwenden müssen, das Verhalten zu antizipieren. Für UX bedeutet dies:

- Konsistente Interaktionsmuster
- Klare Regeln, wann Automatisierung greift
- Begrenzung nicht-deterministischer Outputs in sicherheitskritischen Kontexten

c) Interventionsmöglichkeiten

Nutzer:innen müssen jederzeit in der Lage sein, das Verhalten der KI zu beeinflussen oder zu stoppen. Dies reicht von *Undo-Funktionen* bis zu physischen Not-Aus-Mechanismen. UX-relevante Faktoren:

- Niedrige Einstiegshürden für Eingriffe (keine komplexen Menüs)
- Mehrstufige Eingriffsmöglichkeiten (Feinsteuerung vs. kompletter Abbruch)
- Sichtbarkeit und Erreichbarkeit der Kontrollfunktionen

d) Rückmeldungen & Erklärungen

Kontrollierbarkeit hängt davon ab, ob Nutzer:innen die Auswirkungen ihrer Eingriffe nachvollziehen können. Effektive Feedback-Mechanismen:

- Sofortige visuelle oder akustische Bestätigung
- Erklärung der Veränderung nach einem Eingriff
- Möglichkeit zur Überprüfung, ob die gewünschte Wirkung eingetreten ist

e) Adaptivität mit Nutzerkontrolle

KI kann sich an das Verhalten und die Präferenzen des Nutzers anpassen, sollte dabei aber stets abschaltbare und *übersteuerbare* Mechanismen bieten. Hier ist der Balanceakt entscheidend: zu viel Anpassung ohne Transparenz kann das Gefühl der Kontrolle untergraben.

Praxisbeispiel: In medizinischen Diagnosesystemen kann Transparenz durch erklärbare Modellentscheidungen (XAI) ergänzt werden, während Vorhersagbarkeit und klare Eingriffsmöglichkeiten verhindern, dass Ärzte blind den KI-Empfehlungen folgen.

Die zuvor beschriebenen Dimensionen der Kontrollierbarkeit bilden den allgemeinen Rahmen dafür, wie Menschen mit KI-Systemen interagieren, sie verstehen und steuern können. Während diese Prinzipien in jedem Anwendungsbereich relevant sind, gewinnt eine spezielle Ausprägung besondere Bedeutung in sicherheitskritischen oder hochregulierten Kontexten: **Human Oversight**

Der [AI Act der Europäischen Union](#) macht Human Oversight zu einer verbindlichen Anforderung für Hochrisiko-KI-Systeme. Dabei wird die

Kontrollierbarkeit konkret auf die Frage zugespitzt, wie Menschen gezielt, informiert und wirksam in den Betrieb einer KI eingreifen können. Human Oversight ist damit keine bloße Zusatzfunktion, sondern ein zentrales UX- und Governance-Element, das technische, rechtliche und psychologische Aspekte der Mensch-KI-Interaktion bündelt.

Human Oversight als spezielle Form der Kontrollierbarkeit

3. Definition & Zielsetzung Human Oversight

Human Oversight bezeichnet die systematisch gestaltete Möglichkeit für Menschen, den Betrieb und die Entscheidungen eines KI-Systems zu überwachen, zu bewerten und bei Bedarf einzugreifen. Im Unterschied zu spontanen oder reaktiven Eingriffen ist Human Oversight als *vorgesehener Bestandteil des Systemdesigns* integriert.

Das Ziel von Human Oversight ist zweifach:

- 1. Sicherheit** - Verhindern oder Abmildern von Schäden, die durch fehlerhafte oder unerwünschte Entscheidungen entstehen könnten.
- 2. Verantwortlichkeit** - Sicherstellen, dass es stets einen nachvollziehbaren, menschlichen Entscheidungsträger gibt, der die letzte Verantwortung für kritische Ergebnisse trägt.

In der Praxis umfasst Human Oversight alle Maßnahmen, die gewährleisten, dass:

- Menschen informiert genug sind, um sinnvolle Eingriffe vorzunehmen.
- Eingriffe rechtzeitig erfolgen können, bevor Schaden entsteht.
- Die KI-Nutzung in einen klaren Governance- und Verantwortungsrahmen eingebettet ist.

Der **EU AI Act** definiert Human Oversight explizit als Anforderung an Hochrisiko-KI-Systeme (z. B. in der Medizin, in der Strafverfolgung oder bei kritischer Infrastruktur). Die zugrunde liegende Annahme: KI-Systeme können Fehler machen oder von Trainingsannahmen abweichen - menschliche Aufsicht reduziert das Risiko, dass diese Fehler unentdeckt und unkontrolliert bleiben.

Aus UX-Perspektive bedeutet Human Oversight nicht nur, dass eine Eingriffsmöglichkeit existiert, sondern dass diese *auffindbar, nutzbar und wirksam* ist. Das Oversight-Design muss gewährleisten, dass Nutzer:innen im richtigen Moment die nötigen Informationen und die passenden Werkzeuge haben, um zu handeln - ohne überfordert oder durch unnötige Eingriffe ermüdet zu werden.

4. Design-Pattern für Human Oversight

Human Oversight kann in der Praxis in unterschiedlichen Formen umgesetzt werden. Diese *Design-Patterns* unterscheiden sich vor allem darin, **wann** und **wie intensiv** der Mensch in den Entscheidungsprozess der KI eingebunden ist. Der EU AI Act nennt explizit Mechanismen, die sicherstellen sollen, dass Menschen den Betrieb der KI überwachen und eingreifen können. In der UX-Gestaltung bedeutet das, diese Mechanismen so zu integrieren, dass sie **sichtbar**, **verständlich** und **bedienbar** sind.

a) Human-in-the-Loop (HITL)

Der Mensch überprüft und bestätigt kritische Entscheidungen vor ihrer Umsetzung.

Vorteil: Maximale Sicherheit, da keine kritische Aktion ohne menschliche Zustimmung ausgeführt wird.

UX-Anforderung:

- Klare Benachrichtigung, wenn eine Entscheidung ansteht
- Kompakte, aber aussagekräftige Erklärung der KI-Empfehlung
- Einfacher Mechanismus zur Zustimmung oder Ablehnung

Beispiel: Radiologisches Diagnosesystem, bei dem Ärzt:innen KI-gestützte Befunde vor Freigabe validieren.

b) Human-out-the-Loop (HOTL)

Der Mensch überwacht den laufenden Prozess und kann bei Bedarf eingreifen, muss es aber nicht proaktiv bei jeder Entscheidung tun.

Vorteil: Effizienter, da die KI autonom arbeitet, bis eine Intervention erforderlich ist.

UX-Anforderung:

- Kontinuierliche Statusanzeigen und Prozessvisualisierungen
- Frühwarnungen bei Anomalien oder Risikoindikatoren
- Sofortige Eingriffsmöglichkeiten mit minimalem Reaktionsweg

Beispiel: Autonomes Fahren, bei dem der Fahrer jederzeit übernehmen kann, wenn das System eine kritische Situation meldet.

c) Human-in-Command (HIC)

Der Mensch definiert die übergeordneten Ziele, Grenzen und Rahmenbedingungen und kann den Betrieb der KI jederzeit stoppen oder neu konfigurieren.

Vorteil: Hohe strategische Kontrolle, auch wenn operative Entscheidungen autonom getroffen werden.

UX-Anforderung:

- Leicht zugängliche Konfigurations- und Abschaltfunktionen
- Transparente Darstellung der aktuellen Systemziele und -grenzen
- Logging und Audit Trails, um getroffene Entscheidungen nachzuvollziehen

Beispiel: Militärische Drohnensteuerung, bei der der Operator Einsatzregeln festlegt und jederzeit den Einsatz beenden kann.

Gestaltungsprinzipien über alle Patterns hinweg

1

Sichtbarkeit: Kontrolloptionen müssen leicht auffindbar und jederzeit zugänglich sein.

2

Zeitkritik: Je geringer die Reaktionszeit, desto direkter und weniger verschachtelt muss der Eingriffspfad sein.

3

Informationsdesign: Nur relevante Informationen anzeigen, um Überforderung und „Alert Fatigue“ zu vermeiden.

4

Vertrauenskalibrierung: Interface-Design muss ein realistisches Bild der KI-Fähigkeiten und -Grenzen vermitteln.

5. UX-Herausforderungen bei Human Oversight

Human Oversight stellt nicht nur technische, sondern vor allem gestalterische Herausforderungen. Selbst wenn Eingriffsmöglichkeiten vorhanden sind, kann ihre Wirksamkeit stark eingeschränkt sein, wenn sie aus UX-Sicht nicht optimal umgesetzt werden. Dabei lassen sich die größten Stolpersteine in drei Hauptkategorien einteilen:

a) Aufmerksamkeitsfalle (*Automation Complacency*)

Wenn KI-Systeme über längere Zeit fehlerfrei oder sogar besser als der Mensch arbeiten, neigen Nutzer:innen dazu, ihre Aufmerksamkeit zu reduzieren.

Folge: Eingriffe erfolgen zu spät oder gar nicht, weil Anomalien nicht mehr aktiv überwacht werden.

UX-Ansatz:

- Periodische aktive Bestätigung der Nutzer:innen einfordern („Are you still there?“-Checks in kritischen Prozessen)
- Adaptive Anzeigen, die bei hohem Risiko die Aufmerksamkeit erhöhen
- Schulung und bewusste Sensibilisierung für seltene, aber kritische Eingriffsfälle

b) Alert Fatigue

Wenn zu viele Warnungen oder Eingriffsaufforderungen erscheinen - insbesondere mit geringer Relevanz - tritt das Gegenteil der beabsichtigten Wirkung ein: Nutzer:innen ignorieren auch wichtige Alarme.

Folge: Kritische Warnungen werden übersehen oder reflexartig weggeklickt.

UX-Ansatz:

- Priorisierung von Alerts nach Schweregrad und Handlungsdringlichkeit
- Zusammenfassung von Informationsmeldungen, um Benachrichtigungsflut zu vermeiden
- Möglichkeit für Nutzer:innen, Alarmempfindlichkeit fein einzustellen

c) Erklärungsformat und Handlungsrelevanz

Selbst wenn eine KI ihre Entscheidungen transparent macht, heißt das nicht automatisch, dass Nutzer:innen diese Informationen verstehen oder anwenden können.

Folge: Oversight wird formal erfüllt, aber praktisch wirkungslos.

UX-Ansatz:

- Nutzung verständlicher, nicht-technischer Sprache für Erklärungen
- Ergänzung durch visuelle Darstellungen (Heatmaps, Diagramme, Ablaufvisualisierungen)
- Kontextbezogene Handlungsoptionen direkt im Erklärungsfenster („Jetzt korrigieren“ statt „Gehe zu Menüpunkt 5“)

Zusatzproblem: Balance zwischen Kontrolle und Autonomie

Zu restriktives Oversight-Design kann die Effizienz der KI untergraben, während zu wenig Kontrolle Risiken erhöht. Die UX-Herausforderung besteht darin, **adaptive Kontrollmodi** zu gestalten, die sich an Kontext, Nutzererfahrung und Risikolage anpassen.

6. Messung und Evaluation von Human Oversight

Damit Human Oversight nicht nur als formale Anforderung existiert, sondern tatsächlich wirksam ist, muss er regelmäßig **gemessen, getestet und optimiert** werden. Aus UX-Sicht umfasst Evaluation sowohl quantitative Leistungsdaten als auch qualitative Nutzererfahrungen.

1. Quantitative Metriken

Diese Metriken erfassen messbare Aspekte der Oversight-Wirksamkeit:

- **Eingriffshäufigkeit:** Wie oft greifen Nutzer:innen in den KI-Betrieb ein?
Aussagekraft: Hohe Eingriffsrraten können auf mangelnde KI-Qualität hinweisen, zu niedrige auf unzureichende Wachsamkeit.
- **Zeit bis zum Eingriff (Reaction Time):** Wie lange dauert es, bis Nutzer:innen auf eine kritische Situation reagieren?
Besonders relevant in sicherheitskritischen Szenarien wie Medizin, Luftfahrt oder Verkehr.
- **Fehlervermeidung durch Eingriff:** Anteil der KI-Fehler, die vor Schadenseintritt erkannt und korrigiert wurden.
- **Erfolgsquote der Intervention:** Prozentsatz der Eingriffe, die den gewünschten Effekt hatten.

2. Qualitative Evaluationsmethoden

Diese Methoden beleuchten die subjektive Wahrnehmung, das Vertrauen und die mentale Arbeitsbelastung der Nutzer:innen:

- **Usability-Tests:** Beobachten, wie einfach Nutzer:innen Oversight-Funktionen finden und nutzen können.
- **Kognitive Walkthroughs:** Schritt-für-Schritt-Analyse, ob Nutzer:innen im kritischen Moment die richtige Aktion wählen.
- **Think-Aloud-Protokolle:** Erfassung des Denkprozesses während der Interaktion, um mentale Modelle zu verstehen.
- **Post-Task-Befragungen:** Bewertung von Verständlichkeit, Sicherheitsempfinden und wahrgenommener Kontrolle.

3. Simulation und Szenariotests

Gerade bei selten auftretenden, aber hochkritischen Situationen sind kontrollierte Tests entscheidend:

- **Fault Injection:** Absichtlich Fehlentscheidungen der KI einbauen, um Eingriffsverhalten zu testen.
- **Time Pressure Scenarios:** Messen, ob Nutzer:innen auch unter Stress rechtzeitig reagieren.
- **Mode Confusion Tests:** Prüfen, ob Nutzer:innen wissen, in welchem Automatisierungsmodus sich das System befindet.

4. Kontinuierliche Optimierung

Evaluation ist kein einmaliger Schritt, sondern Teil eines *iterative Design Loops*:

1. Messen
2. Analysieren
3. Interface anpassen
4. Erneut testen

Gerade in adaptiven KI-Systemen kann sich das Nutzerverhalten im Laufe der Zeit ändern, was regelmäßige Re-Evaluationen nötig macht.

06

Mentale Modelle

Kursübersicht > Gestaltungsziele für menschzentrierte KI

Dieses Kapitel behandelt werden Mentale Modelle betrachtet und wie sie bestimmen, wie Menschen verstehen, was eine KI tut und wie sie mit ihr interagieren.

1. Einleitung: Was sind mentale Modelle?

Mentale Modelle sind **innere, vereinfachte Repräsentationen** davon, wie ein System funktioniert, welche Ziele es verfolgt und wie es auf bestimmte Eingaben reagiert. Zum Beispiel definiert Johnson-Laird (1983) mentale Modelle als:

„.... an inner replica of a situation or set of relations, constructed from perception, imagination, or discourse. These models, like a physical model of the solar system or a diagram, represent the structure of the world and are manipulated for reasoning, inference, and understanding, rather than relying on formal logical rules.“

Mit anderen Worten, mentale Modelle sind mentale Repräsentationen - Werkzeuge, um mentale Objekte zu manipulieren, um so Lösungen für Probleme zu finden. Sie entstehen aus Erfahrungen, Beobachtungen und bereitgestellten Erklärungen. In der Mensch-KI-Interaktion dienen sie als kognitive Grundlage, um:

- Systemverhalten vorherzusagen
- Entscheidungen über Eingriffe oder Kooperation zu treffen
- Vertrauen und Arbeitsverteilung sinnvoll zu gestalten

Ein präzises mentales Modell **unterstützt effiziente Zusammenarbeit** und **angemessene Kontrolle**. Ist das Modell jedoch unvollständig oder falsch, kann es zu Missverständnissen, Fehlentscheidungen oder ineffizienter Nutzung führen.

Bspw. haben wir eine mentale Vorstellung davon, was passiert, wenn wir ein Auto starten oder ein Computer ein Programm ausführt. Wir können das Fahrzeug lenken oder Dateien in einem Programm bearbeiten. Ein vollständiges technisches Verständnis aller beteiligten Komponenten oder des dahinterliegenden Codes ist dafür nicht notwendig.

2. Was bedeutet Mental Model Complementary (MMC)?

Mensch und KI besitzen oft unterschiedliche, aber sich ergänzende Wissens- und Verständnisbereiche. Die KI bringt statistische Mustererkennung und Verarbeitungsgeschwindigkeit ein, während der Mensch Kontextwissen, ethische Abwägung und kreative Problemlösung beisteuert. Ziel ist eine **optimale Überlappung**, damit Wissenslücken wechselseitig kompensiert werden.

Der Begriff **Mental Model Complementary (MMC)** beschreibt die Idee, dass die **mentalnen Modelle von Mensch und KI** sich wechselseitig **ergänzen** sollen. Ziel ist ein **gemeinsames Verständnis**, das eine effektive Kooperation ermöglicht.

MMC bedeutet daher:

„Nicht Gleichheit der Modelle, sondern Komplementarität ihrer Stärken.“

3. Warum ist MMC wichtig?

In kollaborativen Mensch-KI-Systemen (z.B. Entscheidungsunterstützung in Medizin, HR, Justiz) kommt es nicht nur auf technische Leistung an, sondern auf ein gutes Zusammenspiel:

- MMC fördert **gegenseitiges Verständnis und interaktive Kontrolle.**
- Sie verbessert die **gemeinsame Fehlerdiagnose.**
- Sie erlaubt, **Verantwortung sinnvoll aufzuteilen.**

Ein Beispiel: Eine KI schlägt eine Diagnose vor, weil sie statistische Zusammenhänge erkennt. Der Arzt ergänzt durch Wissen über seltene Nebenerkrankungen und Patientenbiografie - das mentale Modell des Arztes **komplementiert** das der KI.

4. Gestaltung von MMC in der Praxis

Damit mentale Modelle komplementär werden, braucht es gezielte Gestaltung:

a) Transparenz & Erklärbarkeit

Systeme sollten ihr Vorgehen **offenlegen**, sodass Nutzer:innen Schlüsse daraus ziehen können.

Beispiel: Feature-Visualisierung oder Konfidenzwerte anzeigen.

b) Unterstütztes Modelllernen

Nutzer:innen sollten durch **Feedback, Visualisierungen oder Simulationen** ein mentales Modell entwickeln können.

KI kann den Menschen auch über **eigene Grenzen informieren** (Meta-Kommunikation).

c) Bidirektionale Anpassung

Nicht nur der Mensch passt sich an das System an - das System kann auch **auf den mentalen Zustand des Menschen reagieren**, z.B. durch adaptive Erklärungen oder Warnhinweise.

d) Gemeinsame Aufgabenstruktur

Interfaces sollten **Aufgaben so aufbereiten**, dass sie menschliche und maschinelle Beiträge sichtbar und kombinierbar machen.

e) Dekompositionale Aufgabenverteilung

Eine zentrale Technik zur Förderung von MMC ist die **dekompositionale Aufgabenstrukturierung**:

- Die Gesamtaufgabe wird in Teilaufgaben zerlegt.
- **Mensch und KI** übernehmen jeweils die Komponenten, in denen sie ihre spezifischen Stärken ausspielen können.
- **Beispiel:** In der medizinischen Diagnose übernimmt die KI das Screening großer Bilddatenmengen, der Mensch interpretiert auffällige Ergebnisse im Kontext individueller Patient:innen.

Vorteil: Dekomposition fördert klare Zuständigkeiten und gegenseitiges Vertrauen - jeder Teilnehmende versteht die Rolle des anderen

5. Herausforderungen bei MMC

1

Modellkonflikte: Mensch und KI kommen zu widersprüchlichen Einschätzungen.

2

Modellunsicherheit: Menschen haben kein stabiles Modell der KI - besonders bei intransparentem Verhalten.

3

Kognitive Überlastung: Zu viele Informationen über die Funktionsweise der KI können überfordern.

4

Missverständnisse: Menschen interpretieren KI-Ausgaben nach ihren eigenen kognitiven Mustern, was zu Fehlurteilen führen kann.

6. Empfehlungen zur Förderung von MMC

1

Erklärungen nutzerzentriert gestalten - z.B. durch kontrastive

Erklärungen: „Warum A statt B?“

2

Modellbildung unterstützen - z.B. durch interaktive

Visualisierungen oder kontrolliertes Experimentieren mit dem System

3

Unterschiede sichtbar machen - etwa durch Darstellung

divergierender Einschätzungen zwischen Mensch und Maschine

4

Training & Reflexion - Nutzer:innen sollten explizit über ihr

mentales Modell nachdenken (z.B. in Schulung oder Feedbacksituationen)

5

Systemverhalten adaptiv gestalten - z.B. mehr Erklärungen

bei erkennbarer Unsicherheit oder falscher Nutzung

7. Fazit: MMC als Zukunftsprinzip kollaborativer KI

MMC verschiebt den Fokus weg von der reinen *Nutzerfreundlichkeit* hin zur **kognitiven Partnerschaft**: Mensch und KI sollen nicht identisch, sondern anschlussfähig denken.

Wenn Mensch und System ihre unterschiedlichen Stärken **wechselseitig nutzbar machen**, entsteht ein leistungsfähiges, robustes und verantwortbares Entscheidungssystem.

„Gute KI ist nicht der bessere Mensch - sondern der bessere Partner.“

07

Fazit

Kursübersicht > Gestaltungsziele für menschzentrierte KI

Künstliche Intelligenz verändert nicht nur, **was** Systeme können, sondern auch, **wie** Menschen mit ihnen interagieren. Gerade weil KI-Systeme zunehmend autonome, erklärungsbedürftige und einflussreiche Entscheidungen treffen, ist es nicht mehr ausreichend, ihre Leistung allein an Genauigkeit oder Effizienz zu messen.

Stattdessen müssen **UX-bezogene Eigenschaften** in den Mittelpunkt rücken, um sicherzustellen, dass KI-Systeme **verständlich und vertrauenswürdig** gestaltet werden.

Die behandelten Aspekte - **Vertrauenswürdigkeit, Transparenz, Erklärbarkeit, Kontrollierbarkeit und mentale Modellbildung** - bilden ein eng verzahntes Set an Qualitätsdimensionen. Sie greifen ineinander, bedingen sich gegenseitig und haben gemeinsam ein Ziel: **angemessene und verantwortbare Mensch-KI-Interaktion** zu ermöglichen.

Nur wenn diese Eigenschaften gezielt und **menschzentriert gestaltet** werden, kann KI-Technologie **verantwortungsvoll in gesellschaftliche Entscheidungsprozesse eingebettet** werden.

08

Quellen

Kursübersicht > Gestaltungsziele für menschzentrierte KI

Literaturverzeichnis

Vertrauenswürdigkeit

- European Commission. (2019). *Ethics guidelines for trustworthy AI*. High-Level Expert Group on Artificial Intelligence. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1).
<https://doi.org/10.1162/99608f92.8cd550d1>
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624–635.
<https://doi.org/10.1145/3442188.3445923>

- Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
https://doi.org/10.1518/hfes.46.1.50_30392
- Madsen, M., & Gregor, S. D. (2000). *Measuring human-computer trust*.
<https://api.semanticscholar.org/CorpusID:18821611>
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3), 709–734. JSTOR. <https://doi.org/10.2307/258792>
- OECD. (2019). *Recommendation of the council on artificial intelligence*.
<https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>
- Perez Alvarez, M., Havens, J., & Winfield, A. (2017). *ETHICALLY ALIGNED DESIGN a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems*.

Transparenz

- Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of The Acm*, 59(2), 56–62.
<https://doi.org/10.1145/2844110>
- Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. *Proceedings of the 23rd international conference on intelligent user interfaces*, 211–223. <https://doi.org/10.1145/3172944.3172961>
- Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.3126971>

- Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18.
<https://doi.org/10.1145/3351095.3372833>

Erklärbare KI (XAI)

- Deck, L., Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). A Critical Survey on Fairness Benefits of Explainable AI. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1579–1595. <https://doi.org/10.1145/3630106.3658990>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
<https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007>
- Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250.
<https://doi.org/10.1145/3531146.3534639>

Kontrollierbarkeit

- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/https://doi.org/10.1016/0005-1098(83)90046-8)

- DIN EN ISO 9241-210. (2020). *DIN EN ISO 9241-210:2011-01, ergonomie der mensch-system-interaktion - teil 210: Prozess zur gestaltung gebrauchstauglicher interaktiver systeme (ISO 9241-210:2010); deutsche fassung EN ISO 9241-210:2010* (DIN EN ISO 9241-210:2011-01). Beuth Verlag GmbH.

[**https://doi.org/10.31030/1728173**](https://doi.org/10.31030/1728173)

- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64.

[**https://doi.org/10.1518/001872095779049543**](https://doi.org/10.1518/001872095779049543)

- Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, 59(1), 5–27.

[**https://doi.org/10.1177/0018720816681350**](https://doi.org/10.1177/0018720816681350)

- European Parliament & Council of the European Union. (2024). *Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)*. [**https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng**](https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng)
- Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, 46(1), 50–80.

[**https://doi.org/10.1518/hfes.46.1.50_30392**](https://doi.org/10.1518/hfes.46.1.50_30392)

- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.

[**https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007**](https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007)

- Nielsen, J. (1994). *Usability engineering*. Morgan Kaufmann.
- Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3), 286–297. [**https://doi.org/10.1109/3468.844354**](https://doi.org/10.1109/3468.844354)

- Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, 37(1), 5–19.

<https://doi.org/10.1518/001872095779049516>

- Sheridan, T. B. (2016). Human-robot interaction: Status and challenges. *Human Factors*, 58(4), 525–532.

<https://doi.org/10.1177/0018720816644364>

- Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC Press.
- Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2021). *Engineering psychology and human performance* (5. Aufl.). Routledge.

<https://doi.org/10.4324/9781003177616>

- Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133), 20180085.

<https://doi.org/10.1098/rsta.2018.0085>

Mentale Modelle

- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1), 2–11.

<https://doi.org/10.1609/hcomp.v7i1.5285>

- Ford, M. (1985). *Language*, 61(4), 897–903. JSTOR.

<https://doi.org/10.2307/414498>

- Klien, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a „team player“ in joint human-agent activity. *IEEE Intelligent Systems*, 19(6), 91–95.
<https://doi.org/10.1109/MIS.2004.74>
- Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. Basic Books.

Modul:

KI-Technologien verstehen

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

01

Einleitung

Kursübersicht > [KI-Technologien verstehen](#)

Einleitung: Warum Daten und Informationsverarbeitung die Grundlage für KI-Verständnis sind

Um KI-Systeme sinnvoll einsetzen und bewerten zu können, müssen wir verstehen, wie sie Informationen aufnehmen, verarbeiten und daraus Entscheidungen oder Antworten generieren. In diesem Modul treten wir daher einen Schritt zurück - bevor wir über konkrete Anwendungen oder Ergebnisse sprechen - und betrachten die Grundlagen: **Wie gelangen Systeme überhaupt an Informationen, und was passiert, wenn sie diese „verstehen“ sollen?**

Wenn wir über Künstliche Intelligenz sprechen, sprechen wir im Kern über **Informationsverarbeitung**. Ein KI-System kann nur so gute Ergebnisse liefern, wie die Informationen, auf die es Zugriff hat, sowie deren Strukturierung, Interpretation und Verknüpfung es zulassen.

Leitgedanke: KI verstehen heißt, Informationsverarbeitung verstehen.

Zielsetzung des Moduls

Im Rahmen dieses Moduls erhalten Sie daher nicht nur einen Überblick über zentrale technische Grundlagen, wie etwa die Aufbereitung und Strukturierung von Daten, verschiedene Lernarten oder die Generierung von Ergebnissen (Output), sondern wir verknüpfen diese Aspekte auch mit aktuellen Forschungserkenntnissen zur Informationsverarbeitung bei Systemen und Menschen als Kooperationspartnern.

Wie in den vorherigen Modulen bereits gezeigt wurde, reicht eine rein technische Perspektive auf KI nicht aus. Systeme agieren nicht im luftleeren Raum, sondern werden von Menschen entwickelt, trainiert und genutzt. Ihre Wirksamkeit hängt also immer auch davon ab, wie gut Menschen und Systeme miteinander interagieren, Informationen austauschen und interpretieren.

Fokus: Gemeinsame Informationsverarbeitung von Mensch und System

In diesem Modul bieten wir daher einen **systematischen Einstieg in das Thema Informationsverarbeitung in KI-Systemen** und in die Rolle, die der Mensch in dieser Kooperation spielt. Ausgangspunkt ist das **Modell der integrierten Informationsverarbeitung (Integrated Information Processing)**, das Technik und menschliches Denken gemeinsam betrachtet.



Dabei beschäftigen wir uns unter anderem mit folgenden Fragen:

- Wie müssen Informationen aufbereitet sein, damit Systeme sie „verstehen“ und verarbeiten können?
- Wie lernen Systeme, welche Informationen relevant sind, und wie treffen sie auf dieser Basis Entscheidungen?
- Und schließlich: Wie generieren KI-Systeme Ergebnisse, die für Menschen nachvollziehbar, verständlich und nutzbar sind?

Dieses Verständnis bildet die Grundlage für alle weiteren Themen in der Modulreihe - von Datenqualität und Bias bis hin zu transparenter und vertrauenswürdiger KI. Denn wer versteht, wie Systeme Informationen verarbeiten, kann besser beurteilen, wann ihre Ergebnisse hilfreich, fehlerhaft oder verzerrt sind - und wie Mensch und KI gemeinsam zu guten Entscheidungen kommen.

Kapitelübersicht

1

Input - Technik

Wie funktioniert der Input in ein KI-System auf technischer Ebene und welche Rolle spielen die verfügbaren Daten dabei?

2

Input - Gestaltung

Wie werden Informationen als Input genutzt? Wie beeinflusst dies die Gestaltung von KI-Systemen?

3

Verarbeitung - Technik

Die Frage wie KI-Systeme Daten verarbeiten wird in diesem Kapitel betrachtet.

4

Verarbeitung - Gestaltung

Wie kann die Verarbeitung von KI-Systemen gestaltet werden, damit einzuordnen ist, ob es so arbeitet wie gewünscht?

5

Output - Technik

Mehr als nur ein Ergebnis: KI-Outputs kritisch verstehen und richtig deuten.

6

Output - Gestaltung

Ein guter KI-Output beantwortet mehr als nur die Frage nach dem Ergebnis. Er erklärt das 'Warum', das 'Wie sicher' und das 'Was wäre wenn' für Nutzervertrauen.

LLMs

Dieses Kapitel erklärt, wie LLMs durch Wortvorhersage plausible Texte erzeugen, warum sie aber nichts wirklich verstehen, deshalb 'halluzinieren' und welche Kriterien bei der Auswahl des richtigen Modells entscheidend sind.

02

Input - Technik

Kursübersicht > [KI-Technologien verstehen](#)

1. Einleitung: Warum es wichtig ist, Daten zu verstehen

Wer mit KI-Systemen arbeitet, arbeitet immer auch mit Daten. Ob ein Chatbot Anfragen von Bürger:innen beantwortet, eine KI eingereichte Anträge prüft oder ein Analyse-Tool soziale Trends erkennen soll - der Ausgangspunkt all dieser Systeme sind Daten. Doch was genau sind eigentlich „Daten“? Und warum ist es so entscheidend, ihre Struktur, Herkunft und Qualität zu verstehen, bevor sie in ein KI-System eingespeist werden?

In gemeinwohlorientierten Organisationen wird KI oft eingesetzt, um Prozesse zu entlasten, Zugänge zu erleichtern oder faire Entscheidungen zu unterstützen. Aber ohne Verständnis dafür, was in ein System hineingeht, bleibt unklar, wie man seine Ergebnisse bewerten oder verbessern kann. **Input verstehen heißt, Daten verstehen.**

2. Was sind Daten und was nicht?

Daten sind zunächst nur Zeichen, Zahlen oder Symbole, die etwas in der Welt abbilden. Erst durch Interpretation, also durch das Einordnen dieser Zeichen in einen Kontext, entsteht Bedeutung.

Vom Zeichen zum Wissen

Zeichen: Einzelne Symbole oder Werte (z.B. „25“, „grün“, „Ja“).



Daten: Zeichen, die systematisch erfasst wurden (z.B. „25 Jahre alt“, „grüne Ampel“, „Ja/Nein Antwort“).



Information: Bedeutung, die sich aus Daten in einem bestimmten Kontext ergibt (z.B. „Die Person ist 25 Jahre alt“).



Wissen: Das Verständnis, wie diese Information einzuordnen ist („Menschen zwischen 18 und 30 gelten hier als junge Erwachsene“).

In der Praxis bedeutet das: Wenn eine KI Informationen zu eingereichten Förderanträgen verarbeitet, arbeitet sie nicht mit Wissen, sondern mit Daten - etwa Texten, Zahlenfeldern oder Kategorien. Das

Wissen, wie diese zu interpretieren sind, liegt beim Menschen oder in den Regeln, mit denen das KI-System trainiert wurde.

3. Wie müssen Daten gestaltet sein, um von KI-Systemen genutzt zu werden?

KI-Systeme können nur so gut arbeiten, wie die Daten es zulassen. Damit ein System Daten verarbeiten kann, müssen diese in einer bestimmten Struktur und einem Format vorliegen.

Datenstrukturen und -formate

Strukturierte Daten: Klar definierte Spalten, Werte und Datentypen (z.B. Tabellen, Datenbanken).

Unstrukturierte Daten: Texte, Bilder, Audiodateien, PDFs - also Informationen ohne feste Ordnung.

Halbstrukturierte Daten: Mischformen wie JSON oder XML, die Strukturmerkmale enthalten, aber flexibel bleiben. Beides sind textbasierte Formate, die Informationen in festen Strukturen speichern: JSON vor allem als Schlüssel-Wert-Paare, XML in verschachtelten Tags.

Viele gemeinwohlorientierte Organisationen arbeiten in ihrem Alltag überwiegend mit unstrukturierten Daten - etwa Antragsdokumenten, Berichten oder selbsterstellten Dokumenten mit Freitextantworten.

Diese unstrukturierten Daten können wertvoll sein, müssen aber in strukturierte oder maschinenlesbare Form gebracht werden, bevor sie als Input für eingesetzte KI-Systeme dienen können.

4. Möglichkeiten der Kategorisierung von Daten

Um zu verstehen, welche Art von Daten man einem KI-System zur Verfügung stellt, ist es hilfreich, Daten nach bestimmten Kriterien zu **kategorisieren**. Diese Kategorisierungen helfen dabei einzuschätzen, ob Daten geeignet sind, welche Form der Aufbereitung sie benötigen und welche Schlussfolgerungen sich später aus ihnen ziehen lassen.

Syntax

Ein erster Aspekt betrifft die **Syntax** - also die formale Struktur der Daten. Syntax beschreibt, in welcher Form ein Wert vorliegt: als Zahl, Text, Kategorie oder Wahr/Falsch-Angabe. Diese Unterscheidung ist entscheidend, weil viele KI-Modelle bestimmte Formate erwarten. Ein Text wie „Ja“ oder „Nein“ muss etwa in 0/1-Werte umgewandelt werden, wenn das System nur mit numerischen Eingaben arbeiten kann.

Erscheinung oder Form

Darüber hinaus spielt die **Erscheinung oder Form** der Daten eine Rolle. Damit ist gemeint, wie die Daten erfasst oder dargestellt sind - etwa als Antwortfeld in einem Formular, als Sensormessung oder als Fließtext in einem Bericht. Diese Form bestimmt häufig, wie leicht oder schwer Daten automatisiert weiterverarbeitet werden können. Während eine standardisierte Eingabemaske klare Werte liefert, sind handgeschriebene Dokumente oder unstrukturierte E-Mails für eine KI nur schwer zu deuten.

Zeitlicher Bezug

Ein weiterer wichtiger Aspekt ist der **zeitliche Bezug** der Daten. Daten sind Momentaufnahmen einer bestimmten Realität, die sich mit der Zeit verändern kann. Angaben zu Einkommensverhältnissen, Bevölkerungsdaten oder Nutzungszahlen können nach einigen Monaten oder Jahren bereits veraltet sein. Wenn eine KI also auf Basis alter Daten trainiert wurde, spiegelt sie möglicherweise eine Realität wider, die so gar nicht mehr existiert.

Skalenniveau

Daten lassen sich außerdem nach ihrem **Skalenniveau** unterscheiden - also danach, wie genau sie messbar sind.

- Nominale Daten (z.B. „Farbe der Karte“) lassen sich nicht in einer Rangfolge bringen.
- Ordinale Daten (z.B. „Zufriedenheit: niedrig - mittel - hoch“) hingegen schon.
- Intervall- und Rationskalen (z.B. Temperatur in °C oder Einkommen in Euro) ermöglichen präzise mathematische Berechnungen.

Das richtige Verständnis dieser Unterschiede ist essenziell, weil sie bestimmen, welche statistischen Verfahren und KI-Modelle überhaupt sinnvoll angewendet werden können.

Datentyp

Schließlich ist auch der **Datentyp** selbst von Bedeutung. Ein Datentyp legt fest, ob eine Information als Zahl, Text, boolescher Wert (wahr/falsch) oder komplexere Struktur vorliegt. Ein scheinbar einfacher Unterschied - etwa zwischen einer Zahl, die als Text gespeichert wurde („15“), und einer echten numerischen Variable - kann bei der Verarbeitung durch eine KI große Auswirkungen haben.

Praxisbezug

In der Praxis ist es hilfreich, diese verschiedenen Kategorien im Blick zu behalten. Wer etwa in einer Organisation arbeitet, die mithilfe von KI eingereichte Förderanträge analysiert, sollte sich fragen:

- Sind die Eingabefelder in den Formularen konsistent aufgebaut (Syntax)?
- Liegen die Anträge in digitaler oder gescannter Form vor (Erscheinung)?
- Beziehen sich die Daten auf aktuelle oder ältere Förderperioden (zeitlicher Bezug)?
- Sind die Bewertungskategorien der Gutachter*innen ordinal oder numerisch (Skalenniveau)?
- Und schließlich: Sind die einzelnen Werte korrekt als Text oder Zahl gespeichert (Datentyp)?

Erst wenn diese Grundlagen verstanden und überprüft sind, kann ein KI-System sinnvoll mit den Daten arbeiten und die Organisation sicherstellen, dass die Ergebnisse nachvollziehbar und belastbar bleiben.

5. Datenqualität

Viele Probleme in KI-Projekten entstehen nicht durch den Algorithmus, sondern durch **mangelhafte Datenqualität**. Gründe können z.B. menschliche **Eingabefehler**, fehlende oder uneinheitliche **Standardisierung** (z. B. unterschiedliche Formate wie Datum 01.02.24 vs. 2024-02-01) oder **Systemumbrüche** (wenn verschiedene Systeme Daten unterschiedlich speichern, z. B. fehlende Vorwahlen beim Wechsel des Systems) sein.

Vier zentrale Qualitätsmerkmale

1

Vollständigkeit: Sind alle notwendigen Informationen vorhanden?

2

Genauigkeit: Sind die Daten korrekt und überprüfbar?

3

Konsistenz: Stimmen Daten innerhalb eines Systems überein (z.B. gleiche Schreibweisen, gleiche Einheiten)?

4

Aktualität: Sind die Daten noch relevant oder bereits veraltet?

Ein Beispiel dazu

Eine Organisation möchte mithilfe von KI prüfen, ob Förderanträge vollständig ausgefüllt sind. Wenn jedoch alte Formulare im Umlauf sind oder Einträge unterschiedlich benannt wurden („Straße“ vs. „Str.“), kann die KI falsche Lücken oder Dubletten erkennen und so zusätzliche Arbeit für Mitarbeitende erzeugen, die diese falschen Positive dann händisch filtern müssen.

6. Beziehungen zwischen Daten: Abhängigkeiten, Korrelationen und Kausalität

KI-Systeme analysieren Daten nicht isoliert, sondern immer in ihren **Beziehungen zueinander**. Diese Beziehungen zu verstehen ist zentral, um beurteilen zu können, **was eine KI tatsächlich erkennt - und was sie nur zu erkennen scheint**.

Zunächst lohnt sich ein Blick auf den Unterschied zwischen **abhängigen** und **unabhängigen Variablen**. Eine unabhängige Variable ist ein Faktor, der andere Werte beeinflussen kann, während eine abhängige Variable das Ergebnis oder die Reaktion auf diesen Einfluss darstellt.

Ein einfaches Beispiel: Wenn eine Organisation untersucht, ob das Einkommen einer Person (unabhängige Variable) beeinflusst, ob sie finanzielle Unterstützung beantragt (abhängige Variable), dann kann eine KI diese Beziehung nur dann korrekt erkennen, wenn beide Variablen klar definiert und sauber erfasst sind.

Solche Zusammenhänge bilden die Grundlage vieler KI-Modelle. Doch wichtig ist: **Eine statistische Beziehung bedeutet nicht automatisch, dass ein echter ursächlicher Zusammenhang besteht.** KI-Systeme identifizieren häufig **Korrelationen**, also gleichzeitige Muster oder Bewegungen in den Daten, ohne zu verstehen, **warum** sie auftreten. Kausalität hingegen beschreibt, dass eine Veränderung in einer Variablen tatsächlich eine Veränderung in einer anderen verursacht.

Ein klassisches Beispiel verdeutlicht das: Wenn eine KI in Daten erkennt, dass in Monaten mit höherem Eisverkauf auch mehr Badeunfälle gemeldet werden, besteht zwar eine Korrelation, aber keine Kausalität. Der eigentliche Grund liegt in einer dritten Variable - dem warmen Wetter, das sowohl den Eisverkauf als auch die Zahl der Badeunfälle beeinflusst.

In gemeinwohlorientierten Projekten kann ein ähnliches Risiko auftreten. Eine KI, die Bürger:innenanfragen auswertet, könnte feststellen, dass bestimmte Stadtteile häufiger Beschwerden einreichen. Ohne Kontext könnte dies fälschlicherweise als „höhere Unzufriedenheit“ interpretiert werden - dabei könnten schlicht **unterschiedliche Kommunikationswege** oder **bessere digitale Zugänge** die Ursache sein.

Wer mit KI arbeitet, sollte daher immer fragen:

- Welche Variablen hängen logisch miteinander zusammen und welche nur zufällig?
- Welche Faktoren könnten im Hintergrund wirken, ohne in den Daten sichtbar zu sein?
- Und wie sicher kann ich sein, dass ein Muster tatsächlich eine Ursache-Wirkung-Beziehung darstellt?

Ein KI-System kann Muster sichtbar machen - aber die Interpretation dieser Muster bleibt menschliche Aufgabe.

7. Bias

Ein weiteres zentrales Thema im Umgang mit Daten für KI-Systeme ist **Bias**, also eine **Verzerrung oder Schieflage in den Daten**. Biases sind nicht immer auf den ersten Blick erkennbar, können aber große Auswirkungen auf die wahrgenommene Fairness, Zuverlässigkeit und Akzeptanz eines KI-Systems haben.

Im Kern entsteht ein Bias dann, wenn die Daten, mit denen ein System trainiert oder gefüttert wird, **nicht die tatsächliche Vielfalt oder Verteilung der Realität widerspiegeln**. Die KI „lernt“ dann ein einseitiges Bild - und reproduziert es bei jeder Entscheidung oder Empfehlung.

Man unterscheidet dabei zwei grundsätzliche Arten, wie ein Bias entstehen kann:

Falsche Abbildung der Realität

Hier sind die Daten schlicht **fehlerhaft, unvollständig oder falsch erhoben**. Vielleicht wurden bestimmte Gruppen gar nicht befragt, Datensätze ungleichmäßig aktualisiert oder Eingabefehler nie korrigiert. Ein Chatbot, der Anfragen von Bürger:innen beantwortet, könnte etwa eine Schieflage aufweisen, wenn die zugrunde liegenden Textbeispiele überwiegend aus einer bestimmten Altersgruppe stammen.

Abbildung einer ungleichen Realität

In diesem Fall spiegeln die Daten die reale Welt korrekt wider - doch diese Welt ist selbst **ungleich oder diskriminierend**. Wenn eine KI beispielsweise historische Personaldaten analysiert, in denen Männer häufiger Führungspositionen innehatten, dann „lernt“ sie diese Ungleichheit mit, selbst wenn niemand sie absichtlich eingebaut hat. Sie läuft so Gefahr, diese Ungleichheit zu reproduzieren.

Bias ist deshalb nicht nur ein technisches, sondern vor allem ein **gesellschaftliches Problem**, das sich in die Technologie einschreibt. In gemeinwohlorientierten Projekten ist der Umgang damit besonders wichtig, weil Entscheidungen hier direkt über **Zugang zu Unterstützung, Sichtbarkeit oder Teilhabe entscheiden können**.

Leitfragen zum Erkennen von Bias

- Wer oder was ist in den Daten **überrepräsentiert**?
- Wer oder was **kommt kaum oder gar nicht vor**?
- Welche historischen oder strukturellen Ungleichheiten könnten sich in den genutzten Daten widerspiegeln?
- Welche Werte oder Annahmen liegen in der Datenerhebung selbst verborgen (z.B. Sprache, Begrifflichkeiten, Klassifikationen)?

Bias lässt sich nie vollständig vermeiden - aber er lässt sich erkennen, benennen und abmildern.

Dazu gehört, die Herkunft und Zusammensetzung der Daten kritisch zu prüfen, verschiedene Perspektiven in die Entwicklung einzubeziehen und den Kontext der Datennutzung offenzulegen. Gerade für Organisationen, die im Dienst des Gemeinwohls arbeiten, ist dies ein entscheidender Schritt, um sicherzustellen, dass KI-Systeme nicht unbeabsichtigt bestehende Ungleichheiten fortschreiben, sondern dazu beitragen, **fairere und inklusivere Entscheidungsprozesse** zu fördern.

03 Input - Gestaltung

Kursübersicht > [KI-Technologien verstehen](#)

1. Einleitung: Methoden integrierter Informationsverarbeitung

Damit KI-Systeme Menschen sinnvoll unterstützen können, müssen beide Seiten dieselbe „Sprache“ sprechen und sich darüber austauschen, was notwendig ist, um eine Aufgabe zu lösen. Während Maschinen Informationen als strukturierte Daten, Gewichte und Wahrscheinlichkeiten verarbeiten, deuten Menschen dieselben Informationen in Bedeutungen, Erfahrungen und Zielen.

Die Kunst besteht darin, diese beiden Arten der Informationsverarbeitung miteinander zu verbinden. Genau hier setzt der Gedanke der integrierten Informationsverarbeitung an: Informationen fließen nicht nur *vom Menschen ins System*, sondern auch *vom System zum Menschen zurück*. Nutzende verstehen dadurch, wie ein System arbeitet, können ihre Eingaben korrigieren, und lernen mit der Zeit, wie sie bessere, passgenauere Informationen bereitstellen.

In diesem Kapitel stellen wir drei Methoden vor, die diese Zusammenarbeit zwischen Mensch und Maschine besonders unterstützen:

1 **Information Disclosure** - das System teilt kontextrelevante Informationen über seine Entscheidungen mit den Nutzer:innen so, dass diese ihren Input anpassen können.

2 **Informationen editieren** - Nutzer:innen können Eingaben in das System verändern und so beobachten, wie sich diese Veränderungen auf das System auswirken.

3 **Zeitverlauf** - das System zeigt, wie sich Informationen und Bewertungen über die Zeit hinweg verändern.

Alle drei Methoden sollen die Informationsverarbeitung so gestalten, dass sie verständlich, nachvollziehbar und auf gegenseitiges Lernen ausgelegt ist, sodass Mensch und System bestmöglichen Input für die weitere Verarbeitung generieren.

2. Information Disclosure

Information Disclosure bedeutet, dass ein KI-System mehr Informationen bereitstellt, als nur das bloße Ergebnis.

Statt lediglich „Wohnung geeignet: Ja“ oder „Score: 0,73“ auszugeben, gibt das System auch Einblick in die Gründe und Sicherheiten seiner Einschätzung und die Faktoren, die zu dieser Entscheidung führen könnten. Dadurch können Nutzende besser nachvollziehen, wie Entscheidungen zustande kommen und ob sie auf soliden Daten beruhen oder Unsicherheiten bestehen.

Das Beispiel „Wohnbrücke e. V.“

Die Organisation *Wohnbrücke e. V.* unterstützt Menschen in Not bei der Wohnungssuche in einer Großstadt.

Um Wohnungsangebote systematisch zu bewerten, nutzt sie ein KI-System, das jeder Immobilie einen Eignungswert zwischen 0 und 1 zuweist.

Ein Angebot in der Lindenstraße 12 erhält etwa den Wert 0,82.

Das System zeigt außerdem an:

- **Hauptfaktoren:** Barrierefreiheit (+0,15), Nähe zu Betreuungseinrichtungen (+0,12), moderate Miete (+0,08).
- **Unsicherheiten:** Kein aktuelles Bildmaterial vorhanden, fehlende Angaben zur Heizungsart.
- **Confidence Score:** 0,76 (relativ hohe Sicherheit).

So erkennen die Mitarbeitenden: Das System bewertet das Objekt positiv, aber mit gewissen Unsicherheiten, die auf fehlende Daten zurückzuführen sind.

Durch solche Offenlegungen wird die Entscheidungslogik greifbarer. Mitarbeitende lernen, welche Merkmale besonders wichtig sind und wann sie die Bewertung besser hinterfragen sollten. So können sie ihren Input in das System bestmöglich anpassen.

Gleichzeitig entsteht eine Transparenz, die Vertrauen schafft, sowohl in das System selbst als auch in die Entscheidungen, die darauf basieren.

Reflexionsfrage

Wann wäre es in Ihrem Arbeitskontext hilfreich, zu sehen, wie sicher sich ein System bei seiner Einschätzung ist?

Vorteile von Information Disclosure

- Erhöht Transparenz und Nachvollziehbarkeit.
- Fördert Vertrauen in KI-gestützte Prozesse.
- Hilft, Unsicherheiten zu erkennen und gezielt zu beheben.
- Unterstützt Lernprozesse bei Nutzenden („Wie denkt das System?“).

Grenzen von Information Disclosure

- Gefahr der Überforderung: Zu viele Zahlen oder Indikatoren können verwirren.
- Missverständnisse möglich: Ein hoher Confidence Score heißt nicht automatisch, dass die Aussage des Systems „richtig“ ist, sondern nur, dass das System sich dahingehend sehr sicher ist.
- Datenschutz und Wettbewerbsinteressen können Offenlegungen einschränken, beispielsweise könnte ein Tool zur Bewertung von Krediten nicht ohne größere Probleme die Daten anderer Nutzer:innen offenlegen, um seine Entscheidung zu verdeutlichen.

3. Informationen editieren

Während Disclosure Transparenz schafft, lädt die Methode des Information Editierens zur aktiven Auseinandersetzung ein. Nutzende können Eingaben verändern, um zu sehen, wie das System reagiert. Diese „Was-wäre-wenn“-Szenarien helfen, ein intuitives Verständnis für die Informationsverarbeitung der KI zu entwickeln.

Editieren bei Wohnbrücke e. V.

Die Mitarbeitenden testen verschiedene Annahmen, um die Logik des Systems besser zu verstehen.

Sie wählen wieder das Objekt in der Lindenstraße 12, das aktuell den Eignungswert 0,82 hat.

Nun verändern sie einzelne Eingaben:

Veränderung	Neuer Eignungswert	Wert-Veränderung
Miete sinkt von 850€ auf 700€	0,88	+0,06
Entfernung zur nächsten Sozialstation steigt von 1km auf 2km	0,75	-0,07
Barrierefreiheit entfernt	0,68	-0,14

Die Ergebnisse machen sichtbar:

- Das System legt großen Wert auf Barrierefreiheit (-0,14 Punkte Verlust).
- Auch Entfernung zu Betreuungseinrichtungen wirkt stark.
- Mietkosten sind relevant, aber mit kleinerem Einfluss.

Mitarbeitende verstehen nun, welche Eingaben kritisch sind und wo das System Schwerpunkte setzt. Das hilft ihnen, Daten gezielter zu prüfen oder neue Objekte realistischer einzuschätzen.

Reflexionsfrage

Wie könnten Sie in Ihrem Projekt mit „Was-wäre-wenn“-Szenarien prüfen, ob Ihr System nachvollziehbar arbeitet?

Vorteile des Editierens von Informationen

- Fördert aktives, exploratives Lernen über Systemverhalten.
- Erlaubt es, Hypothesen zu prüfen („Was, wenn das Objekt kleiner wäre?“).
- Macht Zusammenhänge greifbar - besonders hilfreich bei komplexen Modellen.

Grenzen und Risiken des Editierens von Informationen

- Nicht jede Veränderung lässt sich eindeutig interpretieren - insbesondere bei stark vernetzten Merkmalen.
- Gefahr der Überanpassung: Nutzer:innen könnten versuchen, Eingaben „zu optimieren“, statt realistische Daten zu liefern.
- Zusätzlicher technischer Aufwand, da das System flexibel auf Eingabeveränderungen reagieren und Veränderungen visualisieren muss.

4. Zeitverlauf - Entscheidungen nachvollziehbar machen

Die dritte Methode erweitert die Perspektive:

Zeitverlauf meint die Möglichkeit, Veränderungen von Eingaben und Ergebnissen über die Zeit zu beobachten.

Damit wird nachvollziehbar, *wie* und *warum* sich Systementscheidungen entwickeln - ein zentrales Element für Vertrauen und Verantwortlichkeit.

Gerade in Organisationen, in denen Daten laufend aktualisiert werden (z.B. Mietspiegel, Infrastruktur, Energiepreise), kann der Zeitverlauf wichtige Hinweise geben.

Zeitverlauf bei Wohnbrücke e. V.

Das Team verfolgt über drei Monate hinweg, wie sich die Bewertung der **Wohnung Lindenstraße 12** entwickelt:

Datum	Änderungen in den Daten	Eignungswert	Bemerkung
01. Februar	Ursprüngliche Bewertung	0,82	Barrierefrei, mittlere Miete
15. Februar	Neue Info: Heizkosten steigen um 30€	0,79	Leichter Abfall
01. März	Zusatzdaten: Nähe zu Schule (0,8km)	0,83	Verbesserung
01. April	Neue Konkurrenzangebote in der Umgebung	0,76	Sinkende relative Attraktivität

Dieser Verlauf zeigt: Das System reagiert auf neue Informationen dynamisch. Für die Organisation bedeutet das Transparenz - sie kann sehen, warum ein Objekt heute anders bewertet wird als vor einem Monat.

Zudem können die Daten genutzt werden, um Trends zu erkennen:
Verändern sich Bewertungen ganzer Stadtteile?
Wie stark beeinflussen steigende Energiekosten die Eignungswerte insgesamt?

Reflexionsfrage

Wie könnte ein Verlaufs- oder Änderungsprotokoll in Ihrem Projekt helfen, Entwicklungen besser zu verstehen oder zu kommunizieren?

Vorteile von Zeitverläufen

- Erhöht Nachvollziehbarkeit: Zeigt, wie Eingaben und Ergebnisse sich über die Zeit entwickeln.
- Stärkt Verantwortlichkeit: Dokumentiert Entscheidungen und Änderungen, sodass nachvollziehbar ist, wer welche Schritte unternommen hat.
- Macht Lerneffekte sichtbar: Zeigt, wie das System auf Feedback oder Veränderungen reagiert und daraus lernt.
- Fördert Vertrauen: Transparente Entwicklungen schaffen Sicherheit und Vertrauen in das System.

Grenzen und Herausforderungen von Zeitverläufen

- Erhöhter Speicher- und Dokumentationsaufwand.
- Datenschutzfragen (Protokolle enthalten oft sensible Informationen).
- Zu viele Details können den Überblick erschweren.

5. Fazit

Diese drei Methoden - Disclosure, Editieren und Zeitverlauf - verdeutlichen, dass Informationsverarbeitung in KI-Systemen kein einseitiger Vorgang ist.

Sie sind Werkzeuge, um den Dialog zwischen Mensch und Maschine zu fördern.

1

Disclosure schafft Verständnis und Transparenz

2

Editieren ermöglicht aktives Lernen und kontrolliertes Experimentieren

3

Zeitverlauf bietet Nachvollziehbarkeit und Reflexion über Zeit

Gemeinsam bilden sie die Grundlage für KI-Systeme, die nicht nur technisch effizient, sondern auch sozial und ethisch handhabbar sind.

Verarbeitung - Technik

Kursübersicht > [KI-Technologien verstehen](#)

Auf technischer Ebene beschreibt „Verarbeiten“, wie ein KI-System Eingabedaten aufnimmt, analysiert und daraus Ergebnisse oder Entscheidungen ableitet. Verschiedene technische Varianten dieser Verarbeitung werden im Folgenden betrachtet.

Überwachtes Lernen

Ein Verfahren des maschinellen Lernens, bei dem ein System anhand von gelabelten Daten trainiert wird, um später neue Eingaben korrekt zu klassifizieren oder Vorhersagen über Zahlenwerte zu treffen.

1. Einführung in das überwachte Lernen

Überwachtes Lernen (engl.: supervised learning) ist eine Methode des maschinellen Lernens, bei der ein System anhand von gelabelten Daten

trainiert wird. Das bedeutet: Für jede Eingabe gibt es ein bekanntes Ergebnis, das als Orientierung dient.

Beispiel:

Stellen Sie sich vor, wir möchten ein System trainieren, handschriftliche Buchstaben zu erkennen. Jede Buchstabendarstellung im Datensatz ist bereits mit dem richtigen Buchstaben **labelt**, also beschriftet. Das System lernt so, die Muster der Buchstaben zu erkennen.

0 7 1 1 4 9 4 3 4 8 2 2 1 8 6 9 0 3 4 0 2 9

0 7 1 1 4 9 4 3 4 8 2 2 1 8 6 9 0 3 4 0 2 9

2. Wichtige Bestandteile des Lernprozesses

Beim überwachten Lernen spielen zwei zentrale Aspekte eine wichtige Rolle: **Labels** (also die Zielwerte) und die **Aufteilung des Datensatzes** in verschiedene Teile.

Labels - die „richtigen Antworten“

Labels sind die Ergebnisse oder Kategorien, die wir unseren Eingaben (den sogenannten Features) zuordnen.

Beispiel Bilderkennung: Wenn wir ein Foto von einem Verkehrsschild in ein System eingeben, dann ist das passende Label z. B. „Stoppschild“ oder „Geschwindigkeitsbegrenzung 50 km/h“. Die Aufgabe des Modells besteht darin, die Eingabe (das Foto) mit dem richtigen Ausgabewert (dem Label) zu verknüpfen.

Einfache Ja/Nein-Fragen: Manchmal ist die Klassifikation binär. Wir möchten ein System trainieren, das entscheidet, ob eine Person für einen Kredit geeignet ist (Ja) oder nicht (Nein). Dazu nutzen wir einen Datensatz mit vielen Finanzinformationen (z. B. Einkommen, Schulden, Zahlungsverhalten). Das System lernt dann, welche dieser Merkmale entscheidend für die Kreditwürdigkeit sind.

Ohne Labels kann das Modell nicht lernen, ob seine Vorhersagen korrekt sind. Sie sind sozusagen die „Lösungsschablonen“, an denen das Modell sein Wissen überprüft.

Aufteilung des Datensatzes - Trainieren und Testen

Damit das Modell nicht nur die vorhandenen Daten auswendig lernt, sondern allgemein gültige Regeln erkennt und Fehler frühzeitig identifiziert werden, wird der Datensatz in verschiedene Teile aufgeteilt:

Trainingsdaten: Dies ist der größte Teil des Datensatzes. Das Modell „sieht“ diese Daten während der Trainingsphase und versucht, Muster darin zu erkennen.

Testdaten: Dieser Teil wird während des Trainings nicht verwendet. Erst wenn das Modell fertig trainiert ist, werden die Testdaten eingesetzt, um zu überprüfen, wie gut das Modell mit neuen, unbekannten Daten zurechtkommt. So können wir feststellen, ob das Modell wirklich gelernt hat oder ob es sich nur die Trainingsbeispiele „gemerkt“ hat.

In vielen Fällen wird zusätzlich noch ein Validierungsdatensatz genutzt, mit dem die Zwischenergebnisse während des Trainings kontrolliert werden.

Beispiel: Bei einem Fruchtbilder-System lernt das Modell an den Trainingsbildern („Apfel“, „Birne“, „Banane“). Anschließend wird mit den Testbildern geprüft, ob es neue Früchte korrekt erkennt.

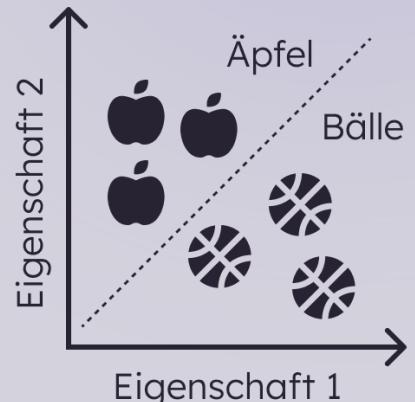
3. Anwendungsbeispiele des überwachten Lernens

Beim überwachten Lernen gibt es zwei zentrale Anwendungsarten: **Klassifikation** und **Regression**.

Klassifikation - Daten in Kategorien einordnen

Bei der Klassifikation werden Eingaben bestimmten Kategorien zugeordnet.

- Beispiel Bilderkennung: Ein Bild wird automatisch die beiden runden Objekte „Äpfel“ oder „Bälle“ erkannt.
- Beispiel Bewertung: Ein System kann Tests oder Aufgaben daraufhin beurteilen, ob sie „vollständig“ oder „unvollständig“ sind.

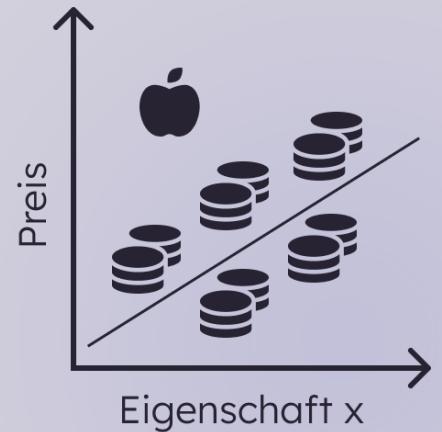


Das Ziel ist es also, qualitative Unterschiede zu erkennen und Daten in klar definierte Gruppen einzuführen. Beim überwachten Lernen werden diese Klassifikationsregeln nicht von Hand aufgeschrieben. Stattdessen

stellt der Mensch eine Reihe von Beispielen mit den richtigen Labels bereit. Das Modell lernt anhand dieser Beispiele, wie es selbstständig auch neue Daten richtig kategorisieren kann. Der Mensch, der die Labels liefert, fungiert dabei gewissermaßen als „Aufsichtsperson“, die den Algorithmus in die richtige Richtung lenkt.

Regression - Vorhersage von Zahlenwerten

Während es bei der Klassifikation um Kategorien geht, beschäftigt sich die **Regression** mit der Vorhersage **kontinuierlicher Zahlenwerte**, auch von Regressionsproblemen. Statt Labels wie „Apfel“ oder „Ball“ wird also eine Zahl vorhergesagt, z. B. ein Preis für Äpfel im zeitlichen Verlauf, die möglichst nahe am realen Wert liegt.



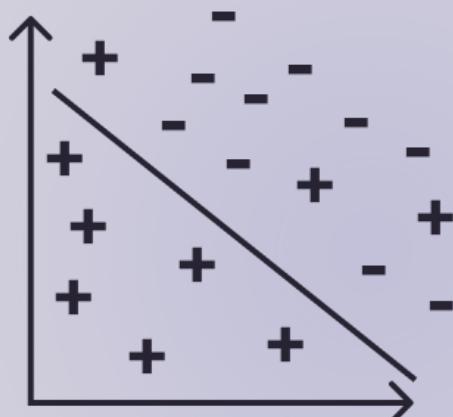
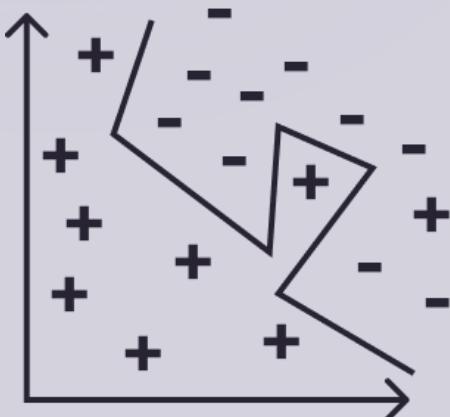
Besonders wichtig ist dabei, dass die Vorhersage meist nicht von einer einzelnen Einflussgröße, sondern vom Zusammenspiel mehrerer Variablen abhängt. Die Regression untersucht also, wie unterschiedliche Faktoren gemeinsam auf die Zielgröße wirken.

Beispiele

- **Einkommen:** Schätzung des Gehalts einer Person auf Basis von Alter, Ausbildung, Berufserfahrung und Branche.
- **Werbung:** Vorhersage der Klickrate einer Online-Anzeige in Abhängigkeit von Text, Gestaltung, Zielgruppe und bisherigen Nutzerverhalten.
- **Verkehr:** Prognose der Anzahl von Unfällen, wobei Faktoren wie Straßenbedingungen, Wetter, Tageszeit und Geschwindigkeitsbegrenzungen einfließen.
- **Immobilien:** Schätzung des Verkaufspreises einer Wohnung oder eines Hauses anhand von Lage, Wohnfläche, Baujahr, Ausstattung und energetischem Zustand.

4. Herausforderungen beim überwachten Lernen

Beim überwachten Lernen gibt es zwei zentrale Probleme, die die Leistungsfähigkeit eines Modells stark beeinflussen können: **Overfitting** und **Underfitting**.



Overfitting

Das Modell lernt die Trainingsdaten zu genau, erkennt keine allgemeinen Muster. Es liefert sehr gute Ergebnisse bei Trainingsdaten, aber schlechte Vorhersagen bei neuen Daten.

Beispiel: Ein Modell wurde nur mit Daten älterer Menschen trainiert und liefert für jüngere Menschen ungenaue Vorhersagen.

Underfitting

Das Modell erkennt die Zusammenhänge in den Daten nicht ausreichend und liefert ungenaue Vorhersagen, auch bei Trainingsdaten.

Beispiel: Eine KI zur Spam-Erkennung, die nur kurze Texte als Spam markiert, ignoriert andere wichtige Merkmale wie Schreibweisen oder Handlungsaufforderungen.

Das Ziel beim überwachten Lernen ist, ein ausgewogenes Modell zu entwickeln, das die richtigen Muster erkennt, ohne sich zu sehr an die Trainingsdaten zu klammern. Nur so kann es auf neue Daten gut generalisieren und verlässliche Vorhersagen treffen.

1

Gelabelte und ausreichende Datenmenge: Damit ein Modell erfolgreich überwacht lernen kann, müssen einige Voraussetzungen erfüllt sein. Zunächst sind **gelabelte Daten** erforderlich, denn ohne bekannte Ergebnisse kann das System nicht lernen. Außerdem ist eine **ausreichende Datenmenge** wichtig, um zu verhindern, dass das Modell unteranpasst und keine sinnvollen Muster erkennt (Underfitting).

2

Qualität und Vielfalt der Daten: Neben der Menge spielen auch die **Datenqualität und Vielfalt** eine entscheidende Rolle. Ein vielfältiger Datensatz hilft, dass das Modell nicht nur die Trainingsdaten auswendig lernt, sondern die zugrunde liegenden Zusammenhänge erkennt und auf neue Daten übertragen kann. So wird Overfitting vermieden. Dabei ist zu beachten, dass natürlich nicht "irgendwelche" Daten genutzt werden sollten. Es geht vielmehr darum, einen möglichst diversen Datensatz zu dem spezifischen Problem zu haben, das das System lösen soll.

3

Schließlich ist eine **klare Zieldefinition** notwendig: Soll das Modell Eingaben klassifizieren, also Kategorien zuordnen, oder numerische Werte vorhersagen, also eine Regression durchführen? Eine präzise Zielsetzung bestimmt den Aufbau des Modells und die Auswahl der geeigneten Daten.

4

Einsatz von Testdaten: Um zu überprüfen, wie gut ein Modell auf neue Daten reagiert, werden **Testdaten** eingesetzt. Diese Daten wurden beim Training nicht verwendet und ermöglichen eine realistische Einschätzung der Leistungsfähigkeit. Zur Bewertung des Modells werden verschiedene **Kennzahlen** herangezogen, wie z.B. Genauigkeit oder Fehlermaße.

Zusammenfassung

Überwachtes Lernen funktioniert nur mit **gelabelten Daten**. Modelle unterscheiden sich je nach Ziel in **Klassifikation**, also der Einordnung in Kategorien, und **Regression**, also der Vorhersage von Zahlenwerten. Ziel ist immer ein ausgewogenes Modell, das weder Overfitting noch Underfitting zeigt. Eine **gute Datenbasis** - qualitativ hochwertig, vielfältig und ausreichend groß - ist entscheidend für den Erfolg eines Modells.

Unüberwachtes Lernen

Beschreibt ein Verfahren des maschinellen Lernens, bei dem ein System ohne vorgegebene Labels in den Daten eigenständig Strukturen, Muster oder Gruppen erkennt.

1. Einführung in das unüberwachte Lernen

Im Gegensatz zum überwachten Lernen arbeitet das **unüberwachte Lernen** (engl.: unsupervised learning) mit Daten, die **keine Labels** enthalten. Das bedeutet: Es gibt keine vorgegebenen Kategorien oder Ergebnisse, an denen sich das System orientieren kann. Stattdessen versucht das Modell, selbstständig Strukturen und Muster in den Daten zu erkennen.

Beispiel für unlabeled Daten sind etwa große Mengen von PDF-Antragsformularen oder Videoaufzeichnungen von Parlamentsdebatten (z.B. auf openparliament.tv). Diese Daten liegen zwar in großer Zahl vor, sind aber nicht vorab in Kategorien eingeteilt oder beschriftet.

Unüberwachtes Lernen kommt immer dann zum Einsatz, **wenn noch keine Kategorien existieren** oder wenn es darum geht, Daten **neu zu ordnen oder in Gruppen einzuteilen**.

Die zentralen Fragestellungen lauten daher:

- Wie lässt sich Struktur in einer großen Menge unlabeled Daten entdecken?
- Wie können diese Daten sinnvoll gruppiert oder zusammengefasst werden?

2. Clustering

Eine der wichtigsten Methoden im unüberwachten Lernen ist das **Clustering**. Darunter versteht man die **Gruppierung von Daten nach ihrer Ähnlichkeit**. Ziel ist es, Datensätze so anzurufen, dass ähnliche Daten in einer Gruppe - einem sogenannten Cluster - zusammengefasst werden.

Beispiel: Stellen Sie sich vor, es liegen viele verschiedene Antragsformulare vor. Ohne vorherige Labels könnte das System sie automatisch in Cluster einteilen, etwa in:

- **Sozialhilfe**
- **Wohngeld**
- **Elterngeld**

So entstehen sinnvolle Gruppen, die eine spätere Analyse oder Verarbeitung erleichtern.

Ein wichtiger Aspekt beim Clustering ist die **Anzahl der Cluster**:

- Mit **mehr Clustern** entstehen feinere Gruppen (hohe Granularität), die Unterschiede sehr detailliert darstellen.
- Mit **weniger Clustern** entstehen gröbere Gruppen (geringe Granularität), die nur die wichtigsten Unterschiede berücksichtigen.

3. Beispiel aus der Praxis

Ein praktisches Beispiel für unüberwachtes Lernen ist die **Auswertung handschriftlicher Dokumente ohne Labels**. Das Ziel besteht darin, verschiedene Buchstaben voneinander zu unterscheiden, ohne dass diese zuvor beschriftet wurden.

Das System geht dabei so vor, dass sie ähnliche Formen automatisch in Gruppen, also **Cluster**, einteilt. So könnten beispielsweise alle „A“s in einem Cluster landen, alle „B“s in einem anderen, und so weiter. Auf diese Weise lassen sich Strukturen in den Daten erkennen, ohne dass vorherige Kennzeichnungen nötig sind.



4. Herausforderungen beim unüberwachten Lernen

Im Gegensatz zum überwachten Lernen bringt das unüberwachte Lernen besondere Schwierigkeiten mit sich. Da **keine Labels** vorhanden sind, gibt es auch keinen klassischen **Test-Datensatz**, mit dem die Ergebnisse überprüft werden könnten.

Das bedeutet: Es gibt keine eindeutig „richtigen“ oder „falschen“ Antworten. Stattdessen zeigt das Modell lediglich mögliche Strukturen auf, die sinnvoll erscheinen können - oder auch nicht. Die **Bewertung der Ergebnisse** ist daher deutlich komplexer und erfordert meist zusätzliches Fachwissen oder weitere Analysen.

Einsatzgebiete: Während überwachte Lernverfahren vor allem dann sinnvoll sind, wenn konkrete Vorhersagen oder Klassifikationen benötigt werden (z.B. Kreditwürdigkeitsprüfung, medizinische Diagnose), eignet sich unüberwachtes Lernen besonders für Exploration, Mustererkennung und Clusterbildung, wenn keine Labels vorliegen und man zunächst Strukturen oder Zusammenhänge in den Daten entdecken möchte.

Zusammenfassung

Unüberwachtes Lernen kommt ohne Labels aus. Sein Ziel ist es, **Strukturen oder Gruppen (Cluster)** in Daten zu erkennen. Die häufigste Methode dabei ist das **Clustering**, bei dem ähnliche Daten automatisch zusammengefasst werden. Allerdings ist die Bewertung der Ergebnisse wesentlich schwieriger als beim überwachten Lernen, da es keine klaren Antworten gibt.

Bestärkendes Lernen

Ein Verfahren des maschinellen Lernens, bei dem ein System durch Versuch und Irrtum lernt und sein Verhalten anhand von Belohnungen oder Bestrafungen schrittweise verbessert, um langfristig den größtmöglichen Erfolg zu erzielen.

1. Einführung in das bestärkende Lernen

Beim **bestärkenden Lernen** (engl.: reinforcement learning) lernt ein KI-System durch **Versuch und Irrtum**. Es erhält **Belohnungen**, wenn es gute Entscheidungen trifft, und **Bestrafungen**, wenn sein Verhalten nicht zielführend ist. Dieser Ansatz ähnelt dem **operanten Konditionieren** in der Psychologie, bei dem Verhalten durch positives oder negatives Feedback geformt wird.

Ein zentrales Merkmal ist, dass es **keine direkten „richtigen“ Lösungen** gibt. Das System lernt stattdessen aus **Feedback, das oft erst nach mehreren Schritten erfolgt**.

Typische Anwendungsbereiche sind Situationen, in denen ein KI-Akteur eigenständig handeln muss, wie etwa:

- **Selbstfahrende Autos**, die erst nach einer Reihe von Entscheidungen Rückmeldung über deren Qualität erhalten.
- **Spiele**, bei denen der Erfolg erst am Ende des Spiels sichtbar wird.

2. Funktionsweise

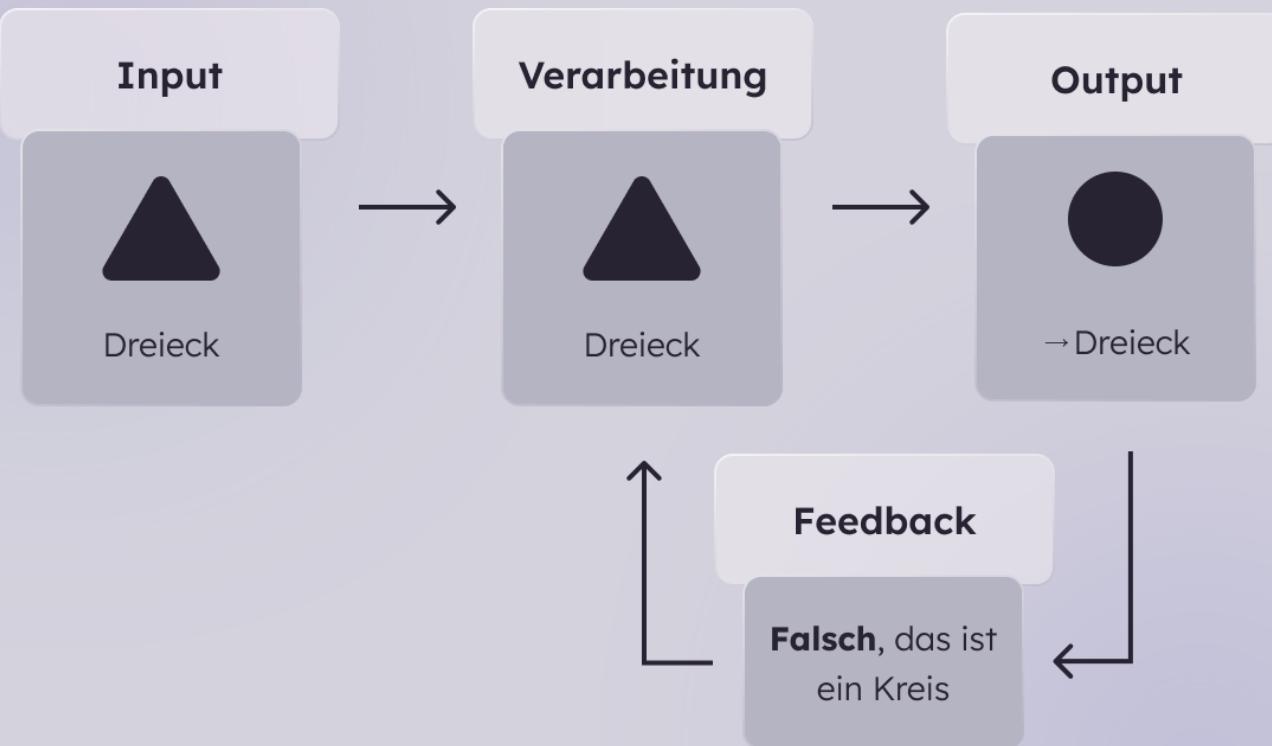
Das System befindet sich in einer **Umgebung**, mit der es ständig interagiert.

- Für jede Aktion erhält es eine Rückmeldung in Form einer Belohnung oder Bestrafung.
- Diese Erfahrungen werden gespeichert und genutzt, um die **Strategie (Policy)** zu verbessern.
- Das Ziel besteht darin, **langfristig die maximale Belohnung** zu erreichen - nicht nur kurzfristig richtige Entscheidungen zu treffen.

3. Besonderheiten

Bestärkendes Lernen unterscheidet sich deutlich vom überwachten und unüberwachten Lernen:

- Es werden **keine großen Datensätze** benötigt, da das System durch direkte Interaktion mit seiner Umgebung lernt.
- Es gibt **keine festen Labels oder Testdaten**.
- Stattdessen basiert der gesamte Lernprozess auf **Erfahrungen**.
- Diese Methode ist besonders geeignet für **dynamische Umgebungen**, in denen sich Situationen ständig ändern und Flexibilität gefragt ist.



Zusammenfassung

Bestärkendes Lernen basiert auf dem Prinzip von **Beloohnung und Bestrafung**. Das System lernt schrittweise aus seinen Erfahrungen und passt sein Verhalten immer weiter an. Es eignet sich besonders für **komplexe, dynamische Umgebungen** wie die Robotik. Das übergeordnete Ziel ist dabei stets die **Maximierung der langfristigen Belohnung**.

05 Verarbeitung - Gestaltung

Kursübersicht > [KI-Technologien verstehen](#)

1. Shapley-Werte - Wer trägt welchen Anteil an der Entscheidung?

Wenn KI-Systeme Entscheidungen treffen, ist das Ergebnis oft nur eine Zahl oder Bewertung. Aber was steckt dahinter? Welche Eingaben haben welchen Anteil daran, dass ein bestimmtes Resultat herauskommt? Genau diese Frage versuchen **Shapley-Werte** zu beantworten - eine Methode, die ursprünglich aus der **Spieltheorie** stammt und heute zu den wichtigsten Werkzeugen gehört, um **Entscheidungsprozesse in KI-Systemen transparent** zu machen.

Vom Spiel zur Entscheidung - die Grundidee

Der Name geht auf den Mathematiker **Lloyd Shapley** zurück, der sich mit fairer Verteilung von Gewinnen in kooperativen Spielen beschäftigte. Stellen wir uns ein Spiel vor, in dem mehrere Personen gemeinsam ein Ziel erreichen - zum Beispiel ein Team, das zusammen einen Gewinn erspielt. Shapleys Frage lautete: *Wie lässt sich dieser Gewinn gerecht auf die Mitspielenden verteilen, abhängig davon, wie stark jede Person zum Gesamterfolg beigetragen hat?*

Diese Idee lässt sich erstaunlich gut auf **KI-Modelle** übertragen:

- Das „**Spiel**“ ist in unserem Fall das Modell, das eine Entscheidung trifft oder eine Vorhersage berechnet.
- Die „**Spieler**“ sind die **Eingabeveriablen**, also beispielsweise Lage, Größe und Kosten einer Wohnung.
- Der „**Gewinn**“ ist die tatsächliche Vorhersage - etwa der Eignungswert eines Gebäudes für eine Wohngruppe - im Vergleich zur durchschnittlichen Bewertung aller untersuchten Objekte.

Das Ziel der Methode: herauszufinden, **wie stark jede einzelne Variable zum Ergebnis beigetragen hat** - und zwar fair, indem alle möglichen Kombinationen von Merkmalen berücksichtigt werden.

Ein praktisches Beispiel

Eine gemeinwohlorientierte Organisation möchte mit Unterstützung eines KI-Systems einschätzen, **welche Immobilien sich für Menschen in Not eignen.**

Das System bewertet verschiedene Objekte anhand von Merkmalen wie:

- Größe der Immobilie
- Entfernung zu einer Betreuungseinrichtung
- Ausstattung und Zustand
- monatliche Mietkosten

Das Modell berechnet für jede Immobilie einen Eignungswert. Eine Wohnung erhält z. B. **nur 0,35 Punkte** (auf einer Skala von 0 bis 1), eine andere **0,75 Punkte**.

Die Verantwortlichen möchten nun verstehen: Warum schneidet die erste so viel schlechter ab?

Mit **Shapley-Werten** lässt sich nachvollziehen, wie stark jedes Merkmal dazu beigetragen hat - etwa:

Merkmal	Shapley-Wert (Beitrag zur Vorhersage)
Größe	+0,15
Lage (Nähe zu Betreuungseinrichtung)	+0,25
Kosten	-0,30
Zustand	-0,05

Solche Werte zeigen: Die höheren Kosten und der schlechte Zustand haben die Bewertung stark gedrückt, während Lage und Größe positiv gewirkt haben. Damit lässt sich nicht nur die Entscheidung des Systems besser verstehen, sondern auch diskutieren, **ob die Gewichtung dieser Faktoren fair und sinnvoll ist.**

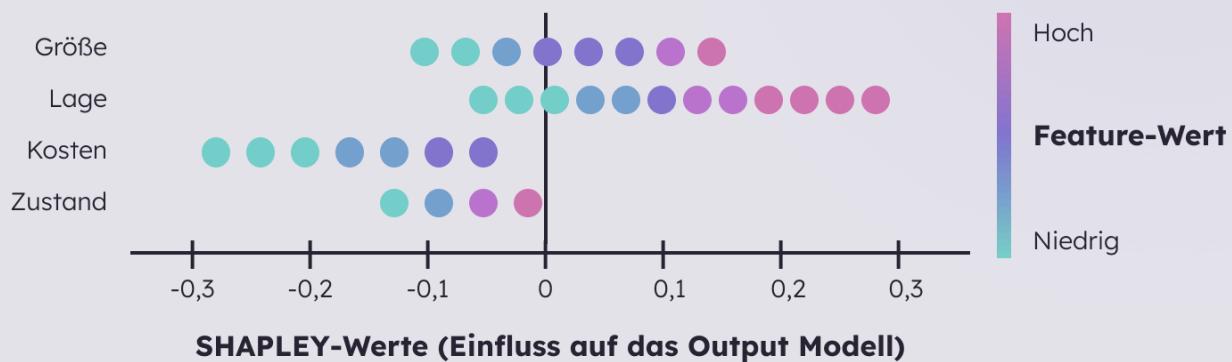
Wie Shapley-Werte berechnet werden

In der Praxis wird für jede Variable berechnet, **wie sich das Ergebnis verändert**, wenn sie zum Modell „hinzugefügt“ oder „weggelassen“ wird - und das über alle möglichen Kombinationen von Variablen hinweg.

Der Durchschnitt dieser Veränderungen ergibt den Shapley-Wert einer Variable.

Da die vollständige Berechnung bei vielen Variablen sehr aufwändig ist, arbeiten Anwendungen meist mit **Näherungsverfahren** oder **Stichproben**.

Viele gängige KI-Frameworks (z. B. SHAP in Python) bieten fertige Implementierungen, die diese Berechnungen automatisiert durchführen. Hier eine beispielhafte Darstellung:



Worauf Sie achten sollten

Die **Interpretation von Shapley-Werten** hängt immer vom **Referenzdatensatz** ab - also den Daten, die als Vergleich herangezogen werden, wenn einzelne Merkmale „fehlen“. Ein unausgewogener oder nicht repräsentativer Datensatz kann zu **verzerrten Ergebnissen** führen. Außerdem gilt:

Shapley-Werte sagen **nichts darüber aus, wie sich das Ergebnis**

verändern würde, wenn ein Merkmal tatsächlich geändert würde. Sie beschreiben nur, welchen Einfluss es im aktuellen Modell hat.

Beispiel: Der Shapley-Wert sagt nicht, „wenn die Wohnung größer wäre, würde sie besser bewertet“, sondern nur, „im aktuellen Datensatz tragen größere Wohnungen im Schnitt positiv zur Bewertung bei“.

Grenzen und Aufwand

Die Methode ist **rechenintensiv**, besonders bei komplexen oder hochdimensionalen Modellen. Für große Sprachmodelle (LLMs) oder neuronale Netze lassen sich daher meist nur **Teilsets von Merkmalen** untersuchen. Dennoch sind Shapley-Werte eines der **robustesten und anerkanntesten Verfahren**, um **Nachvollziehbarkeit und Fairness** in KI-Entscheidungen zu fördern.

2. Partial Dependence Plots

Während **Shapley-Werte** uns zeigen, *welchen Anteil* einzelne Faktoren an einer Entscheidung haben, helfen **Partial Dependence Plots (PDPs)** dabei, *wie genau* diese Faktoren den Ausgang eines Modells beeinflussen.

Mit anderen Worten: Während Shapley-Werte erklären, **wer wie stark mitspielt**, zeigen PDPs, **wie das Zusammenspiel aussieht**.

Was zeigt ein PDP?

Ein **Partial Dependence Plot** stellt visuell dar, wie sich der vorhergesagte Wert eines Modells verändert, wenn sich ein oder mehrere Eingabefaktoren verändern - und alle anderen Faktoren konstant gehalten werden.

Dadurch lässt sich nachvollziehen, welche Beziehung zwischen einem Merkmal und dem Ergebnis besteht:

- Steigt der Wert des Merkmals → steigt oder fällt dann auch die Bewertung durch das System?
- Gibt es Schwellenwerte, ab denen sich der Effekt ändert?
- Wie wirken zwei Merkmale in Kombination aufeinander?

Beispiel: Immobilienbewertung für eine gemeinwohlorientierte Organisation

Kommen wir noch einmal zu unserem Beispiel zurück:

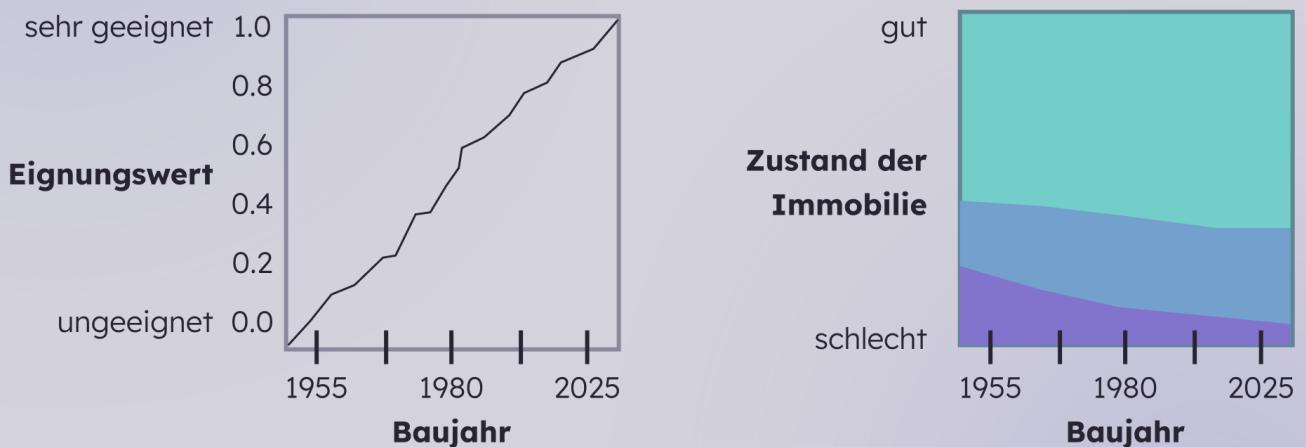
Eine Organisation nutzt ein KI-System, um zu beurteilen, welche Immobilien sich für Menschen in Not eignen. Das System analysiert Merkmale wie Größe, Mietkosten, Baujahr und Zustand der Immobilie und gibt am Ende einen Eignungswert zwischen 0 (ungeeignet) und 1 (sehr geeignet) aus.

Ein **Partial Dependence Plot** könnte nun etwa zeigen, wie das Baujahr der Immobilie den Eignungswert beeinflusst.

In der Abbildung verläuft beispielhaft eine aufsteigende Kurve:

Je neuer das Gebäude, desto höher die Bewertung durch das System.

Das Modell scheint also neuere Immobilien systematisch positiver zu bewerten.



Eine zweite, erweiterte beispielhafte Darstellung zeigt zusätzlich den **Zustand der Immobilie** als zweiten Faktor in einer farbigen **Heatmap**. Darin erkennen wir, dass der Effekt des Baujahrs **abhängig von der Qualität** ist:

- Bei **sehr guter Qualität** spielt das Alter kaum noch eine Rolle - selbst ältere Gebäude erhalten hohe Bewertungen.
- Bei **schlechter Qualität** verstärkt sich dagegen der Effekt des Baujahres deutlich - alte und schlecht erhaltene Gebäude werden klar abgewertet.

Solche Visualisierungen helfen, **Zusammenhänge intuitiv zu erkennen** - auch ohne tief in die Modelllogik einzusteigen.

Wie ein PDP funktioniert

Ein PDP wird erstellt, indem man **den Wert eines Merkmals systematisch variiert** (z. B. das Baujahr von 1950 bis 2020) und dabei beobachtet, **wie sich die Modellvorhersage verändert**, während alle anderen Merkmale konstant bleiben.

Diese Veränderungen werden dann **als Kurve oder Fläche** visualisiert.

- **1D-PDPs** zeigen die Auswirkung **eines einzelnen Faktors** (z.B. Baujahr).
- **2D-PDPs** zeigen die **Wechselwirkung zweier Faktoren** (z.B. Baujahr und Zustand).

Vorteile und Grenzen von PDPs

Vorteile

- Sie sind **intuitiv und leicht verständlich**.
- Sie helfen, **nichtlineare Zusammenhänge** zu erkennen (z.B. Schwellen oder Sättigungseffekte).
- Sie unterstützen Teams dabei, **Modellentscheidungen visuell zu prüfen** - auch ohne technisches Detailwissen.

Grenzen

- PDPs setzen voraus, dass **Faktoren unabhängig voneinander** betrachtet werden können.
In der Realität sind Variablen aber oft **korreliert** - etwa, dass neuere Gebäude meistens teurer sind.
Dadurch können im PDP **unrealistische Kombinationen** entstehen, z.B. „sehr altes Gebäude mit extrem hohem Zustand“, die im echten Datensatz gar nicht vorkommen.
Das Modell zeigt dann zwar eine scheinbare Abhängigkeit, die **in der Praxis aber keine Bedeutung** hat.
- Zudem können PDPs **nur ein oder zwei Merkmale gleichzeitig** darstellen. Für komplexere Interaktionen braucht es andere Verfahren (z.B. Shapley- oder ICE-Plots).

Beispiel für eine Fehlinterpretation

Angenommen, ein PDP zeigt, dass die Bewertung einer Immobilie **stark mit steigender Wohnfläche** zunimmt.

Gleichzeitig sind in den Daten aber größere Wohnungen fast immer **auch teurer**.

Der Plot könnte dann fälschlicherweise suggerieren, dass nur die Größe entscheidend ist, obwohl in Wahrheit die **Mietkosten** der eigentliche Treiber sind.

Solche Effekte zu erkennen, ist Teil einer reflektierten Anwendung dieser Methode - und genau deshalb sind PDPs besonders nützlich, **wenn sie gemeinsam mit anderen Verfahren** wie Shapley-Werten eingesetzt werden.

3. Permutation Feature Importance

Wenn wir verstehen wollen, welche Eingabefaktoren für die Entscheidungen eines KI-Systems tatsächlich wichtig sind, reicht es nicht immer, nur zu wissen, *wie* sie wirken. Oft möchten wir auch wissen, *wie stark* sich das Wegfallen oder Verfälschen eines Faktors oder dessen Reihenfolge im System auf das Ergebnis auswirkt.

Eine Methode, die genau das messbar macht, ist die Permutation Feature Importance (PFI) - zu Deutsch: die Bedeutung eines Merkmals durch Zufallsdurchmischung.

Grundidee: Was passiert, wenn wir die Reihenfolge verändern?

Die Idee hinter der Permutation Feature Importance ist erstaunlich einfach und intuitiv:

Wenn ein bestimmtes Merkmal für die Vorhersage eines Modells wichtig ist, sollte das Ergebnis schlechter werden, sobald wir die Werte dieses Merkmals zufällig durcheinander mischen.

Dadurch wird die ursprüngliche Beziehung zwischen diesem Merkmal und der Zielgröße zerstört.

Je stärker sich der Vorhersagefehler dadurch erhöht, desto wichtiger ist dieses Merkmal für das Modell.

Ein Merkmal, dessen Vertauschung keine oder kaum Veränderungen bewirkt, hat dagegen wenig Einfluss auf die Entscheidungen des Systems - das Modell „ignoriert“ es in gewisser Weise.

Beispiel: Bewertung von Wohnraum für Menschen in Not

Bleiben wir bei unserem laufenden Beispiel.

Das KI-System bewertet Immobilien danach, wie gut sie sich als Wohnraum für Menschen in Not eignen. Eingabefaktoren sind unter anderem:

- Wohnfläche
- Mietkosten
- Entfernung zur Betreuungseinrichtung
- Baujahr
- Ausstattung (Zustand, Barrierefreiheit, etc.)

Nun wollen die Verantwortlichen verstehen, welche dieser Merkmale die Entscheidung des Systems am stärksten beeinflussen.

Dazu wird folgende Vorgehensweise angewandt:

1

Das Modell sagt zunächst mit den echten Daten die Eignungswerte der Immobilien vorher.

Anschließend wird eine Spalte (z.B. das Baujahr) zufällig durchmischt.

2

Damit verliert das Modell die echte Verbindung zwischen Baujahr und Bewertung.

3

Das Modell erstellt erneut Vorhersagen - diesmal mit den „vertauschten“ Werten.

4

Nun wird gemessen, wie stark die Vorhersage an Genauigkeit verliert.

5

Dieser Unterschied wird für jedes Merkmal berechnet, oft mehrfach wiederholt und gemittelt.

Das Ergebnis zeigt, wie sehr das Modell auf jedes Merkmal angewiesen ist.

Beispielsweise könnte sich herausstellen:

Merkmal	Anstieg des Fehlers (in %)	Bedeutung
Mietkosten	+22%	Sehr wichtig
Entfernung zur Betreuungseinrichtung	+15%	Wichtig
Baujahr	+7%	Mittelwichtig
Barrierefreiheit	+4%	Gering
Klimaanlage vorhanden	+1%	Unwichtig

In diesem Fall zeigt sich:

Das Alter der Immobilie hat einen deutlich stärkeren Einfluss auf die Bewertung als etwa die Ausstattung mit Klimaanlage. Das System gewichtet also manche Merkmale stark, andere kaum.

Der technische Ablauf in Kürze

1

Vorhersage mit Originaldaten: Das Modell schätzt, wie geeignet jede Immobilie ist.

2

Permutation eines Merkmals: Die Werte eines Faktors (z. B. Baujahr) werden zufällig neu angeordnet.

3

Vorhersage mit permutierten Daten: Das Modell trifft erneut Entscheidungen, nun ohne die echte Beziehung zwischen Baujahr und Bewertung.

4

Vergleich der Fehler: Wie sehr hat sich der Vorhersagefehler erhöht?

5

Wiederholung: Die Schritte werden mehrfach wiederholt und gemittelt, um zufällige Schwankungen auszugleichen.

Das Ergebnis: ein Wichtigkeitswert pro Merkmal, der zeigt, welche Faktoren das Modell wirklich nutzt, um seine Entscheidungen zu treffen.

Warum das relevant ist

PFI ist besonders dann nützlich, wenn Sie wissen wollen, ob Ihr System auf die richtigen Dinge schaut.

Wenn etwa das Modell in unserem Beispiel feststellt, dass die Postleitzahl oder der Mietpreis besonders großen Einfluss hat, könnte das ein Hinweis auf versteckte soziale oder geografische Verzerrungen (Bias) sein.

Die Methode kann also helfen, Fairness-Probleme frühzeitig zu erkennen und zu adressieren.

Zudem berücksichtigt PFI automatisch auch **Wechselwirkungen zwischen Variablen**:

Wenn sich zwei Merkmale gegenseitig beeinflussen (z. B. Baujahr und Zustand), wird dieser Effekt mitgemessen - denn durch das Durchmischen werden alle Abhängigkeiten zwischen diesem Merkmal und den anderen gleichzeitig aufgehoben.

Vorteile und Grenzen

Vorteile

- **Einfach und modellunabhängig:** Sie funktioniert mit fastem jedem KI-Modell, ohne dass es neu trainiert werden muss.
- **Erhöht Transparenz und Fairness:** Zeigt auf, welche Faktoren in der Praxis tatsächlich Einfluss nehmen.
- **Berücksichtigt Interaktionen:** Auch Kombinationseffekte zwischen Variablen fließen mit ein.
- **Verständlich für Nicht-Expert:innen:** Das Prinzip „wir mischen und schauen, was passiert“ ist leicht nachvollziehbar.

Grenzen und Herausforderungen

- **Kein Verständnis der Richtung:** PFI zeigt nur, *wie stark* ein Faktor wirkt - nicht *ob er positiv oder negativ* wirkt.
- **Zufälligkeit und Streuung:** Da die Methode auf Zufallsdurchmischung basiert, können Ergebnisse schwanken. Eine Mehrfach-Wiederholung (und Mittelung) stabilisiert die Ergebnisse, kostet aber **mehr Rechenzeit**.
- **Keine Aussage über Ursache und Wirkung:** PFI misst Bedeutung, nicht Kausalität. Ein hoher Wert bedeutet nicht, dass dieses Merkmal *verursacht*, dass das Ergebnis so ausfällt - nur, dass es eng damit verknüpft ist.

Beispielhafte Fehlinterpretation

Wenn die Organisation feststellt, dass das Merkmal „Postleitzahl“ sehr wichtig für die Bewertung ist, bedeutet das **nicht**, dass die Lage per se problematisch ist. Es kann sein, dass in bestimmten Stadtteilen schlicht häufiger Immobilien mit schlechter Ausstattung vorkommen – und das Modell diesen Zusammenhang gelernt hat.

PFI hilft also, solche **versteckten Muster sichtbar zu machen**, erfordert aber immer **eine menschliche Interpretation**, um Fehlschlüsse zu vermeiden.

06 Output - Technik

Kursübersicht > [KI-Technologien verstehen](#)

Einleitung: Warum der Output entscheidend ist

Der Output eines KI-Systems ist das sichtbare Ergebnis aller vorhergehenden Verarbeitungsschritte - und damit die Grundlage, auf der Menschen und andere Systeme weiterarbeiten. Er entscheidet darüber, wie nützlich, verständlich und anschlussfähig ein System im konkreten Anwendungskontext ist.

Besonders in gemeinwohlorientierten Organisationen, in denen Entscheidungen oft soziale Folgen haben, ist es wichtig, nicht nur das Ergebnis selbst, sondern auch dessen Art, Herkunft und Aussagekraft zu verstehen. Denn nicht jeder Output ist gleich: Systeme können Texte bewerten, Wahrscheinlichkeiten berechnen, Prognosen abgeben oder sogar neue Daten erzeugen.

Dieses Kapitel gibt einen Überblick über die wichtigsten Output-Formen von KI-Systemen und erläutert, wie sie gelesen, interpretiert und kritisch hinterfragt werden können.

1. Kategorische Outputs

Viele KI-Systeme ordnen Daten in Kategorien ein. Diese Form des Outputs findet sich häufig bei Klassifikationsaufgaben - etwa, wenn ein System E-Mails als „Spam“ oder „Nicht“-Spam“ markiert, oder wenn ein Textanalysetool die Stimmung eines Textes als „positiv“, „neutral“ oder „negativ“ einstuft.

Für gemeinwohlorientierte Organisationen können solche Modelle zum Beispiel eingesetzt werden, um eingehende Anträge zu sortieren oder Texte nach Themen zu gruppieren. Wichtig ist dabei zu verstehen, dass eine Kategorie nicht immer eindeutig „richtig“ ist: ein analysierter Text kann sowohl sachlich als auch emotional gefärbt sein und somit zwei Kategorien zugeordnet werden.

Ein zentrales Merkmal kategorialer Outputs ist die Wahrscheinlichkeit, mit der eine Zuordnung vorgenommen wird. Ein Modell kann etwa schätzen, dass eine Nachricht mit 70 % Wahrscheinlichkeit „positiv“ ist, mit 20 % „neutral“ und mit 10 % „negativ“.

Beispiel aus der Praxis

Eine Organisation, die Bürgeranfragen automatisch vorsortieren möchte, nutzt ein KI-Modell, das E-Mails in die Kategorien „Lob“, „Beschwerde“, „Antrag“ und „Sonstiges“ einteilt. Eine Nachricht wird als „Antrag“ klassifiziert, mit einer Wahrscheinlichkeit von 55 %. Auf den ersten Blick mag das ausreichend erscheinen - doch die zweitwahrscheinlichste Kategorie „Beschwerde“ liegt bei 40 %. Es wäre also riskant, die E-Mail automatisch einem Bearbeitungsprozess zuzuweisen, ohne diesen Unsicherheitsbereich zu berücksichtigen.

Merksatz: Kategorische Outputs sollten nie als absolute Wahrheiten interpretiert werden. Ein Blick auf die zweit- oder drittwahrscheinlichste Kategorie kann helfen, Fehlentscheidungen zu vermeiden.

2. Pattern Matching

Unter *Pattern Matching* versteht man das Erkennen wiederkehrender Muster in Daten. Dabei sucht ein KI-System nach regelmäßigen Abfolgen, Beziehungen oder Ähnlichkeiten zwischen Datenpunkten.

Diese Methode wird vor allem dort eingesetzt, wo es um zeitliche oder sequentielle Zusammenhänge geht - etwa in der Analyse von Verlaufsmustern, Ereignisfolgen oder Textstrukturen.

Beispiel aus der Praxis

Ein gemeinnütziges Gesundheitsprojekt analysiert Gesprächsverläufe aus einer Online-Beratung. Das System erkennt, dass Anfragen, in denen Wörter wie „überfordert“, „allein“ oder „nicht mehr weiter“ vorkommen, häufig in einer Eskalation enden, wenn innerhalb von 24 Stunden keine Antwort erfolgt. Diese Mustererkennung hilft der Organisation, Prioritäten zu setzen und gefährdete Fälle schneller zu identifizieren.

Pattern Matching liefert also keine Bewertung, sondern erkennt Strukturen, die menschlichen Entscheidenden Hinweise geben. Dabei gilt: Muster sind immer statistisch - sie zeigen Wahrscheinlichkeiten, keine Notwendigkeiten.

Reflexionsfrage: Welche Risiken könnten entstehen, wenn eine Organisation erkannte Muster als feste Regeln interpretiert?

3. Numerische Prädiktion

Numerische Prädiktionen gehören zu den wichtigsten Outputs vieler KI-Systeme. Statt Kategorien liefert das Modell hier **Zahlenwerte**, die als Bewertung, Wahrscheinlichkeit oder Score dienen.

Ziel ist es, eine mathematische Funktion zu finden, die auf Basis der Eingabedaten einen quantitativen Output erzeugt - zum Beispiel den geschätzten Wert einer Immobilie, die Wahrscheinlichkeit eines Ereignisses oder einen Prioritätsscore.

Beispiel aus der Praxis

Eine soziale Einrichtung möchte geeigneten Wohnraum für Familien in Not finden. Das KI-Modell bewertet 100 verfügbare Wohnungen anhand von Größe, Lage, Zustand und Mietkosten.

Der Output sieht vereinfacht so aus:

Wohnung	KI-Bewertung (Score 0-100)
A	82
B	76
C	41
D	59

Je höher der Wert, desto besser die Eignung. Die Organisation kann diese Bewertung als Orientierungshilfe nutzen - sollte aber immer prüfen, **welche Merkmale den Score am stärksten beeinflusst haben** (z. B. über Methoden wie Shapley-Werte oder Permutationsanalysen).

Numerische Prädiktionen ermöglichen es auch, die **Abweichung** eines Werts zu quantifizieren - ein Vorteil, wenn Systeme durch Lernen schrittweise optimiert werden sollen.

Merksatz: Numerische Outputs machen Unterschiede messbar - aber nicht automatisch erklärbar. Transparente Modelle helfen, Zahlen richtig einzuordnen.

4. Synthetische Ergebnisse

Synthetische Outputs entstehen, wenn ein KI-System **neue Daten erzeugt**, anstatt vorhandene zu bewerten. Dazu gehören automatisch generierte Texte, Bilder, Musik oder Simulationen.

Im gemeinwohlorientierten Bereich kann diese Art des Outputs beispielsweise genutzt werden, um **Situationen zu simulieren oder alternative Szenarien zu prüfen**.

Beispiel aus der Praxis

Ein Stadtentwicklungsprojekt möchte ermitteln, wie sich neue Grünflächen auf die Lebensqualität in einem Viertel auswirken könnten. Das KI-System erzeugt auf Basis vorhandener Umweltdaten und Bürgerbefragungen synthetische Szenarien, die verschiedene Kombinationen von Bebauungsdichte, Verkehrsaufkommen und Grünanteil zeigen. Diese Simulationen helfen, Entscheidungen über die Stadtplanung zu unterstützen, ohne reale Eingriffe vornehmen zu müssen.

Synthetische Ergebnisse bieten also große Chancen für Planung, Simulation und Bildung. Gleichzeitig stellt sich die Frage nach **ethischer Verantwortung**: Je realistischer synthetische Daten sind, desto größer ist das Risiko, dass sie mit echten verwechselt oder missbräuchlich verwendet werden.

Merksatz: Synthetische Daten sind Werkzeuge zur Exploration -
keine Abbilder der Realität.

5. Forecasting

Forecasting ist die Vorhersage zukünftiger Entwicklungen auf Basis vergangener und aktueller Daten. Anders als bei numerischen Prädiktionen liegt hier der Fokus auf Trends über Zeiträume hinweg.

Beispiel aus der Praxis

Eine Organisation, die Lebensmittelpenden koordiniert, nutzt Forecasting, um den künftigen Bedarf an bestimmten Produkten zu planen.

Das Modell zeigt:

- Wenn die Temperaturen im Winter unter 0°C fallen, steigt die Nachfrage nach warmen Mahlzeiten um durchschnittlich 18%.
- In Ferienzeiten sinkt die Spendenbereitschaft um rund 12%.

Diese Informationen ermöglichen es, Ressourcen effizienter einzuplanen und Engpässe frühzeitig zu vermeiden.

Forecasting erlaubt so die Antizipation von Bedarfen und Risiken - ist jedoch immer von der Qualität der zugrundeliegenden Daten abhängig.

Unerwartete Ereignisse (z. B. Pandemien, politische Krisen) können die Genauigkeit solcher Vorhersagen erheblich beeinträchtigen.

6. Metadaten: Wie gut ist der Output?

Neben den inhaltlichen Ergebnissen liefern viele KI-Systeme sogenannte Metadaten - also Informationen über die Güte ihrer eigenen Entscheidungen.

Zu den wichtigsten gehören Accuracy, Precision und Recall.

Accuracy

Wie viele Vorhersagen des Systems waren insgesamt korrekt?

Beispiel: Von 100 Anträgen erkennt ein System 70 korrekt →
Accuracy = 70%.

Precision

Wie viele der als „positiv“ eingestuften Fälle waren tatsächlich positiv?

Beispiel: 50 Anträge wurden als „dringend“ markiert, aber nur 40 waren es tatsächlich → Precision = $40/50 = 80\%$.

Recall

Wie viele der tatsächlich positiven Fälle wurden erkannt?

Beispiel: Es gab 60 wirklich dringende Anträge, 40 davon wurden richtig erkannt → $\text{Recall} = 40/60 = 66,7\%$.

Bedeutung in der Praxis

Eine Organisation, die Anträge nach Dringlichkeit sortiert, sollte dann auf einen hohen **Recall** achten, wenn das Übersehen eines Falls (*False Negative*) gravierende Folgen hätte.

Beispiel: Ein Sozialamt prüft Notfallhilfen für obdachlose Personen. Wenn das System einen wirklich dringenden Antrag übersieht, erhält jemand in akuter Not keine schnelle Hilfe. Deshalb ist es wichtiger, möglichst alle echten Notfälle zu erkennen, auch wenn einige weniger dringende Anträge fälschlicherweise als dringend markiert werden.

Wenn hingegen Falschalarme (*False Positives*) problematisch sind – etwa weil sie Ressourcen binden – ist eine **hohe Precision** wichtiger. Beispiel: Dieselbe Organisation hat nur begrenzte Notfallbetten. Wenn zu viele nicht dringende Fälle fälschlich als dringend markiert werden, könnten echte Notfälle keinen Platz bekommen. Hier ist es entscheidend, dass fast alle als dringend eingestuften Fälle tatsächlich dringend sind.

Reflexionsfragen

Fragen, die man sich im Rahmen des Outputs stellen könnte:

1

Welche Form von Output produziert das KI-System, mit dem Sie arbeiten (z.B. Textklassifikation, Score, Simulation)?

2

Wie könnte die Darstellung der Ergebnisse verbessert werden, um sie für die Zielgruppe verständlicher oder nützlicher zu machen?

3

Wie stark würden Sie sich auf die Ergebnisse verlassen, wenn das System zusätzlich seine Accuracy oder Confidence mitliefert?

Fazit

Das Verständnis verschiedener Output-Formen ist entscheidend, um KI-Systeme sinnvoll in gemeinwohlorientierten Kontexten zu nutzen. Ob kategoriale Zuordnung, numerische Prädiktion, Forecasting oder Simulation - der Output ist immer nur so gut wie seine Interpretation. Die Herausforderung liegt darin, Ergebnisse nicht als absolute Wahrheiten, sondern als Hilfsmittel zur Entscheidungsunterstützung zu begreifen. Nur dann kann KI in gemeinwohlorientierten Organisationen das leisten, was sie soll: Prozesse verbessern, ohne Verantwortung zu ersetzen.

Output - Integrierte Informationsverar- beitung

Kursübersicht > KI-Technologien verstehen

Outputs von KI-Systemen sind die Ergebnisse, die ein KI-Modell produziert, nachdem es Eingabedaten verarbeitet hat. Beispiele: Vorhersagen, Klassifikationen, Wahrscheinlichkeiten, Texte, Bilder etc.
Kurz: Was das Modell am Ende „ausspuckt“.

Was ist bei Interaktion mit KI-Systemen im Bezug auf den Output relevant?

Wenn Nutzer:innen mit einem KI-System interagieren, möchten sie oft verstehen:

- **Warum** hat die KI diese Entscheidung getroffen? → *Local Feature Relevance*
- **Wie sicher** ist sich die KI in ihrer Antwort? → *Confidence Estimation*
- **Was könnte anders sein**, damit das Ergebnis anders ausfällt? → *Counterfactual Explanation*

1. Local Feature Relevance (Lokale Merkmalsbedeutung)

Hier wird **erklärt, welche Eingabemerkmale für eine einzelne spezifische Vorhersage wichtig waren.**

Beispiel bei einer medizinischen Diagnose-KI:

„Für diese eine Vorhersage waren Alter und Blutdruck besonders einflussreich, Geschlecht dagegen weniger.“

Kurz: Welche Faktoren haben in genau diesem Fall das Ergebnis bestimmt?

Typische Anwendung im Interface: Heatmaps, Balkendiagramme, Tooltipps „Dieses Merkmal hatte den größten Einfluss“

Beispiel mit Diabetes-Diagnose

Eine Ärztin gibt die Patientendaten in die KI ein. Das Modell gibt eine Vorhersage aus: „**Der Patient hat ein hohes Risiko, in den nächsten Jahren Diabetes zu entwickeln.**“

Die Local Feature Relevance zeigt für diesen einen Patienten:

Merkmal	Einfluss auf die KI-Vorhersage
Langfristiger Blutzuckerwert (HbA1c)	sehr Hoch
BMI (Übergewicht)	hoch
Alter	moderat
Bewegung pro Woche	gering
Geschlecht	kaum Einfluss

Interpretation: Die Ärztin kann jetzt sehen, dass der hohe HbA1c-Wert und das Übergewicht die Hauptgründe für die hohe Risiko-Vorhersage sind. Das bedeutet nicht, dass die KI „hohe Wahrscheinlichkeit für hohe Wahrscheinlichkeit“ ausgibt, sondern dass die KI das **konkrete Risiko für diesen Patienten** als hoch einschätzt und erklärt, warum.

2. Confidence Estimation (Konfidenzschätzung)

Das ist die **Einschätzung des Modells, wie sicher es sich bei seiner eigenen Vorhersage ist**. Meist wird dies als Wahrscheinlichkeit oder Score ausgegeben.

Wichtig: Hohe Confidence bedeutet nicht automatisch, dass die Vorhersage korrekt ist - nur, dass das Modell „glaubt“, dass sie korrekt ist.

Kurz: Wie sicher ist das Modell bei seiner Antwort?

Typische Anwendung im Interface: Wahrscheinlichkeitsanzeigen, Farbcodierung (z.B. grün = sicher, rot = unsicher)

Beispiel mit Diabetes-Diagnose

Nach der Risiko-Prognose zeigt das System:

Diabetes-Risiko: 78% Wahrscheinlichkeit

Vertrauensniveau des Modells: 92%

Die Ärztin erkennt dadurch:

- Die KI ist **sehr sicher**, obwohl die Entscheidung komplex ist.
- Sie kann die Informationen vorsichtig weitergeben: „Die Daten deuten stark auf ein erhöhtes Risiko hin.“
- Wäre die **Confidence z.B. nur 40%**, wäre Vorsicht angesagt. Eventuell müssten weitere Tests gemacht werden.

3. Counterfactual Explanation (Was-wäre-wenn-Erklärung)

Eine counterfaktische Erklärung zeigt, welche minimale Veränderung an der Eingabe nötig wäre, damit das Modell zu einem anderen Ergebnis kommt.

Beispiel:

„Die Kreditbewerbung wurde abgelehnt. Hätte ihr Einkommen 400€ höher gelegen, wäre sie angenommen worden.“

Kurz: Welche kleine Änderung hätte das Ergebnis verändert?

Typische Anwendung im Interface:

Interaktive „Was-wäre-wenn“-Slider: Nutzer:innen können Werte verändern und sieht direkt, wie sich das Ergebnis ändert.

Beispiel mit Diabetes-Diagnose

Die KI gibt zusätzlich eine counterfaktische Empfehlung:

„ Wenn der BMI um 2 Punkte reduziert wird oder der Patient 2 zusätzliche Sporneinheiten pro Woche durchführt, sinkt das Diabetes-Risiko von 78% auf 45%. “

Im Interface kann die Ärztin z.B. einen Schieberegler bewegen:

- BMI von 31 → 29 → Risiko sinkt sichtbar
- Bewegung 1h/Woche → 3h/Woche → Risiko sinkt weiter

Diese Erklärung ist **handlungsorientiert**, weil sie nicht nur zeigt, was das Risiko ist, sondern auch, was man konkret tun könnte.

UX-Beispiel: Kreditbewilligung

Stellen Sie sich ein Dashboard für Kreditentscheidungen vor:

1

Local Feature Relevance → Balken zeigt: „Einkommen +400€, Schulden -100€ → stärkster Einfluss auf Entscheidung“

2

Confidence Estimation → Ampel/Prozentangabe: „KI ist zu 85% sicher, dass der Antrag abgelehnt wird“

3

Counterfactual Explanation → Interaktives Widget: „Wenn Sie ihr Einkommen um 400€ erhöhen oder Schulden reduzieren, würde der Antrag genehmigt werden“

Resultat für UX:

- Nutzer:innen verstehen die Entscheidung (Local Feature Relevance)
- Nutzer:innen wissen, wie zuverlässig die KI ist (Confidence Estimation)
- Nutzer:innen können potenzielle Maßnahmen ausprobieren (Counterfactual Explanation)

Fazit

Damit Nutzer:innen die Ergebnisse eines KI-Systems verstehen und ihnen vertrauen können, reicht der reine Output (z. B. ein Risiko-Wert oder eine Entscheidung) nicht aus. Erst durch ergänzende Erklärungsmechanismen wird der Output **interpretierbar und nutzbar**.

- **Local Feature Relevance** zeigt, **warum** das Modell zu genau diesem Ergebnis kam.
- **Confidence Estimation** macht transparent, **wie sicher** das Modell in seiner Vorhersage ist.
- **Counterfactual Explanation** eröffnet **konkrete Handlungsmöglichkeiten**, indem sie zeigt, wie das Ergebnis durch kleine Änderungen beeinflusst werden kann.

In der UX führt die Kombination dieser drei Aspekte zu **Vertrauen, Nachvollziehbarkeit und Kontrolle**: zentrale Voraussetzungen für verantwortungsbewusste und gebrauchstaugliche KI-Systeme.

08

LLMs

[Kursübersicht](#) > [KI-Technologien verstehen](#)

LLMs sind große Sprachmodelle, die vorhersagen, wie ein Text fortgesetzt werden könnte, basierend auf einer riesigen Textmenge.

1. Einführung: Was sind Large Language Models (LLMs)?

Large Language Models (LLMs) sind KI-Systeme, die auf der Grundlage großer Textmengen trainiert werden, um Sprache zu verstehen, zu verarbeiten und selbstständig Texte zu erzeugen. Sie bilden die Grundlage vieler moderner Anwendungen wie Chatbots, automatische Übersetzungen oder Textanalysen.

Das folgende Video gibt einen kurzen Überblick über die Funktionsweise von LLMs sowie typische Anwendungsszenarien und deren Grenzen. Diese Aspekte werden in den nächsten Abschnitten vertieft.



<https://youtu.be/KZ5LL1xhAg4>

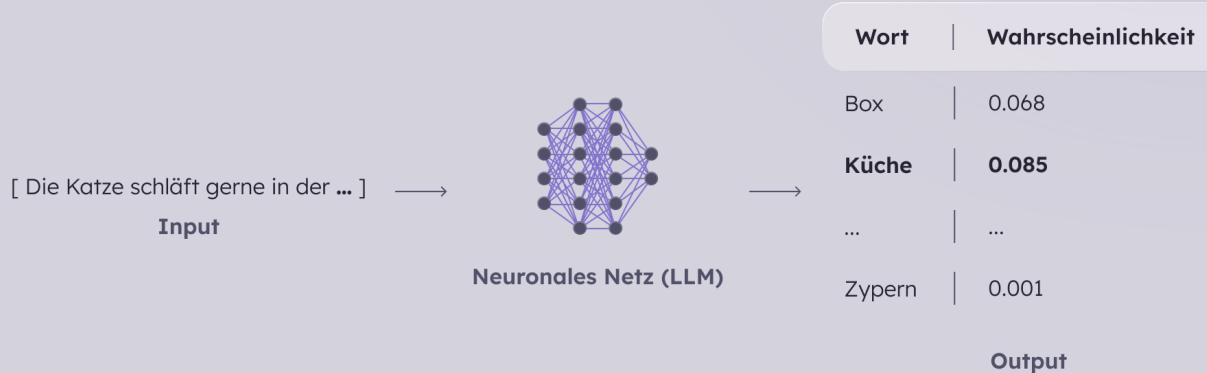
2. Funktionsweise von Large Language Models (LLMs)

Ein **Large Language Model (LLM)** lernt, Sprache zu verstehen und selbstständig Texte zu erzeugen. Im Kern basiert es auf einer einfachen, aber wirkungsvollen Idee:

Es wird darauf trainiert: das nächste Wort in einem Satz vorherzusagen.

Zum Beispiel:

[Die Katze schläft gerne in der ...] → Welches Wort kommt als nächstes?



Dazu benötigt das Modell eine enorme Menge an Textdaten - etwa aus dem Internet, aus Büchern, Artikeln oder anderen Quellen. Diese Daten müssen **nicht manuell beschriftet** werden, da das Modell selbst lernt, Sprachmuster zu erkennen.

Im Training entdeckt das LLM typische **Muster und Strukturen** der Sprache, beispielsweise:

- Welche Wörter häufig zusammen vorkommen
- Wie unterschiedliche Sätze aufgebaut sind
- Welche Bedeutungen und Zusammenhänge zwischen Wörtern bestehen

Sobald das Modell trainiert ist, kann es **neue Texte generieren**, Wort für Wort - oder genauer gesagt, **Token für Token**. Aber was genau ist ein **Token**?

Tokens - die Bausteine der Sprache

Ein **Token** ist eine kleine Einheit von Sprache, die das Modell verarbeitet. Das kann ein ganzes Wort, ein Wortteil oder sogar ein Satzzeichen sein.

Beispiele:

- Das Wort „*Haus*“ ist ein Token
- Das Wort „*Häuserbau*“ könnte in zwei Tokens zerlegt werden:
„*Häuser*“ und „*bau*“

LLMs verarbeiten Sprache anders als Menschen also **nicht in ganzen Sätzen oder Silben**, sondern in diesen kleineren Einheiten. Durch diese Tokenisierung wird die Sprache sehr flexibel und kann detaillierter verwendet werden.

Kreativität und Variation bei der Texterzeugung

Wenn das LLM ein neues Wort (oder Token) vorhersagt, wählt es nicht immer die **wahrscheinlichste** Option. Stattdessen kann es aus mehreren **guten Möglichkeiten** auswählen. Das sorgt für **abwechslungsreiche und kreative Texte**. Deshalb können zwei Antworten auf dieselbe Frage leicht unterschiedlich ausfallen.

Kann KI verstehen? Das Gedankenexperiment des Chinesischen Zimmers

Ein bekanntes Gedankenexperiment, das hilft, die Grenzen von LLMs zu verstehen, ist das „**Chinesische Zimmer**“ des Philosophen **John Searle** (1980).

Stellen Sie sich vor, eine Person sitzt in einem geschlossenen Raum. Sie versteht **kein Chinesisch**, hat aber ein Handbuch mit **Regeln**, wie sie auf chinesische Zeichen richtig mit anderen Zeichen reagieren kann. Durch diese Regeln kann sie auf Fragen in chinesischer Schrift **korrekte Antworten** geben. So könnten Außenstehende denken, dass die Person **Chinesisch** versteht.

Tatsächlich folgt sie aber nur **formalen Anweisungen**, ohne den **Bedeutungsinhalt** der Sprache zu begreifen.

Searle nutzte dieses Gedankenexperiment, um zu zeigen:

Auch wenn ein Computer (oder ein LLM) scheinbar intelligente Antworten gibt, bedeutet das nicht, dass er wirklich versteht, was er sagt. Das Modell verarbeitet nur Symbole nach Regeln, ähnlich wie die Person im chinesischen Zimmer.

Bezug zu LLMs

LLMs funktionieren ganz ähnlich: Sie erkennen Muster in Sprache und erzeugen darauf basierend plausibel klingende Texte.

Doch sie verstehen keine Bedeutungen im menschlichen Sinn. Sie haben kein Bewusstsein, keine Absichten und keine eigenen Gedanken.

Das „Chinesische Zimmer“ regt dazu an, über menschliches und maschinelles Verstehen nachzudenken. Searle stellt die Idee in den Raum, dass ein System zwar auf sprachliche Eingaben sinnvoll reagieren kann, aber ohne dabei tatsächlich zu *verstehen*, was es sagt. Im Kontext moderner LLMs wird diese Frage erneut relevant: Wenn ein Modell Texte analysiert und Antworten generiert, zeigt es dann Intelligenz und Verstehen oder lediglich die Fähigkeit, sprachliche Muster zu erkennen und zu reproduzieren?

Der Chinese Room fordert uns also heraus, die Grenze zwischen echter Erkenntnis und bloßer Symbolverarbeitung von Systemen kritisch zu hinterfragen.

Reinforcement Learning from Human Feedback (RLHF)

Nach dem Grundtraining wird das Modell oft noch durch ein Verfahren namens **Reinforcement Learning from Human Feedback (RLHF)** bzw. bestärkendes Lernen verfeinert.

Dabei bewerten Menschen die Antworten des Modells, zum Beispiel danach,

- wie hilfreich,
- verständlich
- oder angemessen eine Antwort ist.

Das System nutzt Rückmeldungen, um sein Verhalten anzupassen und Vorhersagen zu verbessern. Dadurch entstehen Antworten, die **sprachlich flüssiger und konsistenter** wirken. Alle modernen LLMs wie ChatGPT oder Claude haben diese Form von menschlicher Feinjustierung durchlaufen, bevor sie auf den Markt gekommen sind. Wir merken also, ganz ohne den Menschen geht es nicht.

3. Grenzen von LLMs

So leistungsfähig Large Language Models auch sind, besteht das Risiko, dass ihre Texte **sprachlich überzeugend und intelligent wirken**, aber **inhaltlich nicht korrekt** sind. LLMs kennen keine absolute Wahrheit; sie erzeugen Inhalte ausschließlich auf Basis von **statistischen Wahrscheinlichkeiten**, die aus Trainingsdaten und Bewertungen abgeleitet werden, und stoßen dabei an folgende **Grenzen und Herausforderungen**:

1. Halluzinationen

LLMs können falsche oder frei erfundene Informationen liefern, die **plausibel klingen**, aber **nicht stimmen**. Das passiert, weil sie keine Fakten prüfen, sondern nur wahrscheinlich klingende Texte erzeugen.

2. Fehlendes Verständnis

LLMs „verstehen“ Inhalte nicht im menschlichen Sinn. Sie wissen nicht, was Wörter *bedeuten*, sondern nur, wie sie typischerweise zusammen vorkommen.

3. Veraltetes Wissen

Wenn ein Modell nicht regelmäßig aktualisiert wird, kennt es keine **aktuellen Ereignisse** oder **neuen Daten** nach dem Zeitpunkt seines Trainings.

4. Bias (Voreingenommenheit)

Da LLMs auf menschlichen Texten trainiert werden, übernehmen sie auch **gesellschaftliche Vorurteile** oder **einseitige Darstellungen**, die in den Daten vorkommen.

5. Datenschutz und Urheberrecht

In Trainingsdaten können geschützte Inhalte enthalten sein, was rechtliche und ethische Fragen aufwirft. Gerade Bild generierende Systeme sehen sich häufig mit Vorwürfen von Urheberrechtsverletzung konfrontiert.

4. LLMs im Vergleich

Es gibt heute mehrere große **Large Language Models**, die von verschiedenen Unternehmen entwickelt wurden. Sie basieren alle auf ähnlichen Prinzipien, unterscheiden sich aber beispielsweise in **Größe, Trainingsdaten, Zugänglichkeit, Fähigkeiten und Zielrichtung**.

Beispiele bekannter LLMs

Modell & Anbieter	Besonderheit / Fokus	Einsatz & Nutzen	Lizenz / Offenheit
GPT-4 / GPT-4o von OpenAI	Sehr leistungsfähig, vielseitig (Text, Code, Analyse, Konversation)	Schreiben, Programmieren, Wissensarbeit, Chatbots	Proprietär (Cloud-basiert)
Gemini von Google DeepMind	Multimodal (Text, Bild, Code, Video), eng mit Google-Ecosystem verknüpft	Multimodale Anwendungen, Such- und Wissensintegration	Proprietär
Claude von Anthropic	Fokus auf Sicherheit, Ethik, transparente KI-Antworten	Sichere, erklärbare KI-Nutzung in sensiblen Bereichen	Proprietär
Llama von Meta AI	Offen zugänglich, stark für Forschung & Fine-Tuning	Eigene Anpassungen, Forschung, interne Nutzung	Teilweise offen (Open-Weight, Lizenzbeschränkungen)
Mistral / Mixtral von Mistral AI (EU)	Europäischer Fokus auf Effizienz, Datenschutz, Open-Source-Ansatz	On-Premises-Lösungen, datenschutzsensible Anwendungen	Offen (Apache 2.0 / Open-Weight)

Offene Modelle (Open Source)

Offene oder „open-weight“ Modelle (z.B. **Llama**, **Mistral**, **Zephyr**) gewinnen stark an Bedeutung. Dabei handelt es sich nicht unbedingt um echte Open-Source-Modelle, denn dafür müssten auch die Trainingsdaten und der gesamte Trainingsprozess offenliegen. Open-Weight-Modelle machen lediglich die Modellgewichte öffentlich, werden aber häufig trotzdem als Open Source bezeichnet. Diese Modelle ermöglichen:

- **Datenhoheit & Datenschutz** (lokaler Betrieb, keine Cloudpflicht)
- **Anpassbarkeit** (Fine-Tuning, eigene Trainingsdaten)
- **Kostenkontrolle & Unabhängigkeit** von US-Plattformen

Aber: eigener Betrieb erfordert **technisches Know-how, Rechenressourcen und Wartung.**

Checkliste: Welches LLM passt zu meinem Projekt?

1. Trainingsdaten: Welche Art von Daten wurde verwendet?

- Offene Internetdaten → breite Allgemeinbildung, viele Sprachstile
- Lizenzierter / kuratierte Daten → verlässlicher, präziser, kontrollierter Inhalt

Beispiel: GPT-4 trainiert auf einer Mischung aus Webdaten, Büchern und Artikeln → gut für gezielte generelle Textgenerierung.

2. Zugänglichkeit: Wie leicht lässt sich das Modell nutzen?

- Kommerziell (z.B. GPT, Claude) → einfach via API nutzbar, Support vorhanden
- Open Source (z.B. Llama, Mistral) → volle Kontrolle, Anpassung möglich, keine Lizenzkosten

Beispiel: Llama 3 kann lokal eingesetzt werden → ideal für Projekte mit Datenschutzanforderungen.

3. Fähigkeiten / Modality: Welche Art von Daten soll verarbeitet werden?

- Text → alle klassischen LLMs
- Multimodal (Text, Bild, Audio) → Gemini, GPT-4 multimodal

Beispiel: Für ein Projekt, das Bildbeschreibungen generieren soll → GPT-4 multimodal oder Gemini.

4. Ziele / Schwerpunkt: Was soll das Modell erreichen?

- Breiter Einsatz / kreative Texte → GPT, Claude
- Effizienz, Datenschutz, leichte Integration → Mistral, Llama

Beispiel: Ein datenschutzfreundlicher interner Chatbot → Mistral oder Llama sind besser geeignet als kommerzielle APIs.

5. Weitere Kriterien (optional)

- **Kosten:** Open-Source oft günstiger, kommerzielle Modell oft nutzungsbasiert
- **Community / Support:** Größere Modelle wie GPT haben mehr Dokumentation und Nutzerfeedback
- **Anpassbarkeit:** Open-Source Modelle können feinjustiert oder in eigene Pipelines integriert werden

5. Fazit

1

Grundprinzip: LLMs lernen, das **nächste Wort vorherzusagen**, indem sie auf riesigen Textmengen Muster erkennen.

2

Tokens: Sie arbeiten mit **Tokens**, den kleinsten Einheiten der Sprache, was ihnen ermöglicht, flexibel und detailliert Texte zu erzeugen.

3

Kreativität und Variation: Durch geschickte Auswahlmechanismen entstehen **abwechslungsreiche und kreative Texte**, selbst bei identischen Eingaben.

4

Feinabstimmung mit RLHF: Reinforcement Learning from Human Feedback verbessert die Antworten durch menschliches Feedback und macht sie nützlicher, verständlicher und sicherer.

5

Grenzen: LLMs **verstehen Inhalte nicht wirklich**, können **Halluzinationen erzeugen** und sind anfällig für Bias. Sie ersetzen menschliches Urteilsvermögen nicht, sondern unterstützen es.

6

Praxisbezug: LLMs sind mächtige Werkzeuge für Textgenerierung, Analyse, Übersetzung und kreative Aufgaben - ihr Potential entfaltet sich besonders, wenn **menschliche Kontrolle und kritische Prüfung** einbezogen werden.

Kurz gesagt: LLMs sind beeindruckende Sprachwerkzeuge, aber **keine denkenden Wesen**. Sie kombinieren **mathematische Mustererkennung** mit menschlicher Anleitung, um Texte zu erzeugen, die sinnvoll, nützlich und kreativ wirken.

09 Quellen

Kursübersicht > [KI-Technologien verstehen](#)

Literaturverzeichnis

- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. <https://arxiv.org/abs/1702.08608>
- eGov-Campus. (2021). *KI in öffentlichen verwaltungen*.
https://learn.egov-campus.org/courses/kiverwaltung_uzl_2021-1/overview
- Molnar, C. (2025). *Interpretable machine learning: a guide for making black box models explainable* (Third edition). Christoph Molnar.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). „Why should i trust you?“ Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.
- Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE access: practical innovations, open solutions*, 8, 42200–42216.

- Schrills, T. P. P. (2025). *Integrating humans and artificial intelligence in diagnostic tasks: Automation-related user experience & interaction in explainable AI / Integration von Mensch und Künstlicher Intelligenz bei diagnostischen Aufgaben: Automatisierungsbezogene User Experience & Interaktion in erklärbarer KI* [Doctoral dissertation, Universität zu Lübeck]. https://epub.uni-luebeck.de/handle/zhb_hl/3417
- Shapley, L. S. & others. (1953). *A value for n-person games*.
- University of Helsinki & MinnaLearn. (2018). *Elements of AI*.
<https://www.elementsofai.com/>

Modul:

05 Automatisierungs- potenziale erkennen

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

01

Einleitung

[Kursübersicht](#) > [Automatisierungspotenziale erkennen](#)

In diesem fünften Modul vertiefen wir das Thema Automatisierung und den gezielten Einsatz von KI-Systemen in gemeinwohlorientierten Organisationen. Während die vorherigen Module vor allem die Bedeutung einer menschenzentrierten Perspektive auf KI, grundlegende Funktionsweisen von KI-Systemen sowie erste Evaluations- und Interpretationsansätze aus dem User Experience Design in den Fokus gestellt haben, richtet sich der Blick nun stärker auf die Frage:

Wann, wie und in welchem Ausmaß sind Automatisierung und KI in konkreten Arbeitsprozessen tatsächlich sinnvoll und verantwortungsvoll?

Wie bereits in den vorherigen Kapiteln gezeigt, reicht es nicht aus, KI-Systeme nur technisch zu betrachten. Ihre Einführung verändert Arbeitsabläufe, Rollen, Verantwortlichkeiten und Entscheidungsstrukturen und betrifft damit immer auch Menschen, Organisationen und Machtverhältnisse.

Ziel dieses Moduls ist es, Ihnen einen **praxisnahen, systematischen Denk- und Entscheidungsprozess** an die Hand zu geben, mit dem Sie

bewusster abwägen können, **ob und wie Automatisierung in Ihrer Organisation eingesetzt werden sollte** - und wann bewusst darauf verzichtet werden sollte.

Dabei geht es ausdrücklich nicht darum, möglichst viele Prozesse zu automatisieren. Vielmehr geht es darum, **passende, verantwortungsvolle und realistische Entscheidungen** zu treffen.

Im Zentrum stehen dabei fünf grundlegende Leitfragen, die Sie durch dieses Modul begleiten:

1

Erfüllt meine Aufgabe überhaupt die Voraussetzungen, um automatisierungstauglich zu sein?

2

Möchten wichtige Stakeholder überhaupt, dass dieser Prozess automatisiert wird?

3

Welcher Automatisierungsgrad ist sinnvoll und welche Konsequenzen hat er für den Arbeitsprozess?

4

Ist ein KI-System oder ein Large Language Model (LLM) eine geeignete Technologie, um das definierte Ziel zu erreichen?

5

Welche weiteren Voraussetzungen müssen geklärt sein, bevor ein Automatisierungsprojekt gestartet wird?

Diese Fragen sind bewusst einfach formuliert, greifen jedoch tief in organisatorische, ethische, technische und soziale Aspekte ein. Sie sollen dazu anregen, nicht vorschnell zu handeln, sondern bewusst zu reflektieren, **wo der tatsächliche Mehrwert von Automatisierung liegt und wo nicht**.

Im folgenden Video wird ein Überblick über die unterschiedlichen Aspekte bei der Identifikation gegeben, auf die in den folgenden Kapiteln näher eingegangen wird.



<https://youtu.be/0GpnGeD7BCM>

Ein durchgehendes Anwendungsbeispiel: Beratung im gemeinwohlorientierten Kontext

Um diese Fragen greifbarer zu machen, arbeitet dieses Modul mit einem durchgehenden, bewusst allgemein gehaltenen Beispiel:

Stellen Sie sich eine gemeinwohlorientierte Organisation vor, die Beratungsangebote für Menschen in schwierigen Lebenssituationen anbietet. Dazu gehören unter anderem:

- psychosoziale Erstberatung
- Unterstützung bei der Orientierung im Hilfesystem
- Information zu staatlichen und zivilgesellschaftlichen Unterstützungsangeboten
- Weitervermittlung an spezialisierte Stellen

Die Organisation wird über Telefon, E-Mail und ein Online-Formular kontaktiert. Monatlich gehen mehrere hundert Anfragen ein. Die Mitarbeitenden müssen jede Anfrage sichten, einordnen, priorisieren und entscheiden, wie weiter vorgegangen wird. Schon heute sind sie stark ausgelastet, und die Nachfrage steigt kontinuierlich.

Vor diesem Hintergrund entsteht die Frage, ob und wie bestimmte Teilprozesse automatisiert werden könnten, zum Beispiel:

- eine erste Kategorisierung der Anfragen
- das Bereitstellen grundlegender Informationen
- die Priorisierung nach Dringlichkeit
- oder die Unterstützung bei der Dokumentation

Dieses Beispiel wird in den folgenden Kapiteln immer wieder aufgegriffen und variiert, um die verschiedenen Überlegungen rund um Automatisierung, KI-Einsatz und Levels of Automation konkret erlebbar zu machen.

Grundgedanke von Automatisierung

Automatisierung wird häufig mit Effizienz, Zeitersparnis und Kostenreduktion in Verbindung gebracht. Gerade in Organisationen mit begrenzten Ressourcen kann das sehr attraktiv erscheinen. Gleichzeitig sind viele Arbeitsprozesse im gemeinwohlorientierten Bereich hochsensibel:

- Sie betreffen Menschen in belastenden, teilweise kritischen Situationen.
- Sie erfordern Empathie, Kontextverständnis und situative Entscheidungen.
- Sie sind eng mit Vertrauen und Beziehung verbunden.

Deshalb gilt hier ganz besonders: Automatisierung ist kein Selbstzweck. Nicht alles, was technisch automatisierbar ist, sollte auch automatisiert werden. Was „Automatisierung“ in diesem Modul bedeutet, erfahren Sie im ersten Unterkapitel.

Aufbau des Moduls

Um Sie Schritt für Schritt durch diese komplexe Fragestellung zu führen, denkt das Modul grob folgende Bereiche ab:

- **Grundlagen: Was bedeutet Automatisierung im gemeinwohlorientierten Kontext?** (inkl. Chancen, Grenzen und typische Missverständnisse)
- **Prozesserkennung und -analyse:** Wie identifizieren Sie einen Prozess, der (vielleicht) automatisiert werden kann?
- **Automatisierungspräferenzen und Stakeholder-Perspektiven:** Wer ist betroffen und wie unterschiedlich können Erwartungen an Automatisierung ausfallen?
- **Levels of Automation:** Welche Abstufungen gibt es und welche Konsequenzen entstehen für Menschen, Organisation und Nutzer:innen?
- **KI und LLMs als Automatisierungstechnologie:** Wann sind sie geeignet - und wann nicht?
- **Praxisorientierte Checkliste und Entscheidungsunterstützung:** Ein konkretes Werkzeug, das Sie in Ihrer Organisation anwenden können.

Dieses Modul soll Sie nicht zu einer bestimmten Entscheidung drängen, sondern Sie in die Lage versetzen, informierte, reflektierte und kontextangepasste Entscheidungen zu treffen - im Sinne Ihrer Organisation und der Menschen, für die Sie arbeiten.

02 Automatisierung verstehen

Kursübersicht > Automatisierungspotenziale erkennen

Bevor konkrete Prozesse in Ihrer Organisation automatisiert oder durch KI unterstützt werden, ist es wichtig, zunächst ein gemeinsames Grundverständnis davon zu entwickeln, **was Automatisierung ist - und was nicht**. In der Praxis werden die Begriffe **Digitalisierung**, **Automatisierung** und **Künstliche Intelligenz (KI)** häufig vermischt oder synonym verwendet. Für fundierte Entscheidungen ist jedoch eine saubere Abgrenzung notwendig.

Unterschied zwischen Digitalisierung, Automatisierung und KI- Einsatz

Digitalisierung

Unter dem Begriff der Digitalisierung werden verschiedene Phänomene zusammengefasst, die im Zusammenhang mit Computern, Datennetzen und digitaler Infrastruktur stehen. Im Kern geht es dabei um die Umwandlung analoger Informationen in digitale Formate. Gleichzeitig umfasst Digitalisierung heute sehr viel mehr als nur das „Einscannen“ von Dokumenten oder die Nutzung digitaler Medien. Sie betrifft auch:

- die Produktion und Verteilung immaterieller Güter
- die Verarbeitung großer Datenmengen
- den Einsatz von Algorithmen
- den Aufbau einer digitalen Infrastruktur (Hardware, Software, Netzwerke, Daten, Standards)

Digitale Technologien wirken dabei längst nicht mehr nur auf Kommunikation und Medien, sondern beeinflussen auch gesellschaftliche, politische und kulturelle Prozesse, und das bereits seit der Nutzung elektronischer Datenverarbeitung in Verwaltungen seit den 1960er- und 1970er-Jahren. Kurz gesagt:

Digitalisierung schafft die technischen und strukturellen Grundlagen, auf denen Automatisierung und KI überhaupt erst möglich werden.

Automatisierung

Automatisierung bezeichnet die Einrichtung und Durchführung von Arbeits- und Produktionsprozessen, so dass der Mensch nicht mehr unmittelbar in jeden einzelnen Schritt eingreifen muss. Prozesse (einschließlich Steuerung, Regelung und teilweise auch Kontrolle) laufen selbstständig ab.

Dabei kann sich Automatisierung auf unterschiedliche Ebenen beziehen:

1

Verfahrensautomatisierung: Einzelne Arbeitsschritte oder Teiltätigkeiten werden automatisiert (z. B. das automatische Versenden von E-Mails nach einer Anmeldung).

2

Prozessautomatisierung: Ein kompletter Prozess wird automatisiert (z. B. Terminvergabe oder Datenübertragung zwischen Systemen).

3

Vollautomatisierung: Ein gesamter Ablauf entfällt für den Menschen fast vollständig (z. B. automatische Entscheidungen ohne menschliches Eingreifen).

Wichtig ist: Automatisierung bedeutet nicht zwangsläufig den Einsatz von KI.

Viele Prozesse lassen sich auch mit einfachen, regelbasierten Systemen automatisieren.

KI-Einsatz

Künstliche Intelligenz kann als eine besondere, datenbasierte Form von Automatisierung verstanden werden. Sie basiert wie in Modul 4 gezeigt unter anderem auf:

- Machine Learning
- Neuronalen Netzen
- Large Language Models (LLMs)

Vorhersagen zu treffen oder Inhalte zu generieren. Viele Aufgaben, die Künstliche Intelligenz ermöglicht es Systemen, Muster zu erkennen, früher ausschließlich von Menschen ausgeführt wurden, können heute teilweise oder vollständig von KI-Systemen übernommen werden.

Dabei ist entscheidend: KI ist eine mögliche Form der Automatisierung, aber nicht immer die beste, einfachste oder sinnvollste Lösung.

In vielen Fällen ist ein klassisches, regelbasiertes System transparenter, günstiger und besser kontrollierbar als ein KI-System.

Warum Automatisierung im gemeinwohlorientierten Bereich besonders sensibel ist

Gerade in gemeinwohlorientierten Organisationen ist der Umgang mit Automatisierung und KI von besonderer Bedeutung. Zwar gibt es hier ähnliche Hürden wie in anderen Organisationen - etwa Datenschutz, Kosten, technische Infrastruktur oder fehlendes Know-how - doch kommen zusätzliche, verstärkende Faktoren hinzu:

Ressourcen

Gemeinwohlorientierte Organisationen verfügen häufig über **begrenzte finanzielle und personelle Ressourcen**. Automatisierung kann hier ein großes Potenzial bieten:

- Zeitersparnis für Mitarbeitende
- Entlastung von Routineaufgaben
- bessere Skalierbarkeit von Angeboten

Gleichzeitig bedeutet dies auch:

- Fehlentscheidungen sind schwerer zu korrigieren
- investierte Mittel lassen sich oft nicht so leicht ersetzen
- falsche Automatisierung kann langfristig mehr Schaden als Nutzen bringen

Daraus entsteht ein Spannungsfeld: Der Druck, neue Technologien nicht zu verpassen, trifft auf die Notwendigkeit, besonders sorgfältig und effizient vorzugehen.

Vertrauen

Vertrauen ist für jede Organisation zentral - für gemeinwohlorientierte Organisationen jedoch existentiell:

- Sie arbeiten häufig mit vulnerablen Gruppen
- Sie sind für viele Menschen eine der wenigen Anlaufstellen
- Sie sind auf die Unterstützung von Ehrenamtlichen angewiesen

Ein schlecht umgesetztes automatisiertes System - beispielsweise ein unpersönlicher oder fehleranfälliger Chatbot - kann dazu führen, dass sich Menschen:

- nicht ernst genommen fühlen
- sich zurückziehen
- das Vertrauen in die Organisation verlieren

Gleichzeitig kann auch intern Vertrauen zerstört werden, wenn Ehrenamtliche und Mitarbeitende das Gefühl haben, ihre Arbeit werde durch Technik entwertet oder unzureichend unterstützt. Ein Verlust von Vertrauen bedroht die grundlegenden Strukturen gemeinwohlorientierter Arbeit.

Beispiel: Gelungene vs. nicht gelungene Automatisierung

Positives Beispiel (teilautomatisiert, unterstützend)

Eine Organisation möchte das Matching von Freiwilligen mit passenden Einsatzstellen verbessern. Eine Software erstellt auf Basis von Interessen, zeitlichen Verfügbarkeiten und Standort Vorschläge. Die finale Entscheidung wird jedoch weiterhin von Mitarbeitenden getroffen. Daraus ergeben sich folgende Vorteile:

- Zeitersparnis im Auswahlprozess
- menschliche Kontrolle bleibt erhalten
- Freiwillige fühlen sich weiterhin individuell wahrgenommen

→ **Effizient, unterstützend, vertrauensfördernd**

Negatives Beispiel (vollautomatisiert, entkoppelt)

Eine Beratungsstelle für Menschen mit psychischen Erkrankungen ersetzt den Erstkontakt vollständig durch einen Chatbot. Die Implementierung ist komplex und teuer. Betroffene fühlen sich nicht ernst genommen, da sie keinen menschlichen Kontakt mehr haben. Gleichzeitig fällt für Mitarbeitende eine wichtige Informationsquelle weg: die direkte Interaktion mit Ratsuchenden. Das erzeugt folgende Probleme:

- Hoher Ressourcenaufwand
- Vertrauensverlust bei Betroffenen
- Verlust wertvoller sozialer Signale

→ **Ineffizient, belastet Vertrauen, geht an Mitarbeitenden vorbei**

Reflexionsfragen für Ihre Organisation

Bevor Sie über Automatisierung oder KI nachdenken, stellen Sie sich bitte folgende Fragen:

- Welche Probleme möchten wir eigentlich lösen?
- Geht es um Zeit, Geld, Qualität oder Entlastung von Menschen?
- Was würde verloren gehen, wenn dieser Prozess automatisiert wird?
- Wer könnte sich ausgeschlossen oder nicht ernst genommen fühlen?
- Welche Rolle spielt Vertrauen in diesem Prozess?

Diese Fragen bilden die Grundlage für das nächste Kapitel, in dem wir genauer betrachten, **ob und in welchem Umfang ein Prozess überhaupt für Automatisierung geeignet ist.**

Eignung der Automatisierung einschätzen

Kursübersicht > Automatisierungspotenziale erkennen

Nachdem wir gesehen haben, was gute und weniger geeignete Anwendungsräume für Automatisierung und KI sind, stellt sich die zentrale Frage:

Woran können Sie festmachen, ob ein Prozess in Ihrer Organisation für Automatisierung geeignet ist - und ob er überhaupt automatisiert werden sollte?

Gerade in einer gemeinwohlorientierten Organisation ist diese Entscheidung besonders sensibel. Es geht nicht nur um Effizienz, sondern auch um Vertrauen, Verantwortung und die Qualität menschlicher Beziehungen.

Eine realistische Einschätzung der Einsatzmöglichkeiten von KI knüpft direkt an die in Modul 4 vorgestellten Grundlagen zu menschzentrierter Systemgestaltung an.

Zur ersten Orientierung können Sie sich an folgenden Leitfragen orientieren:

- Ist der Prozess regelmäßig und wiederkehrend oder nur gelegentlich relevant?
- Gibt es klare Eingaben und erwartbare Ausgaben?
- Ist der Prozess stark datenbasiert oder stark von menschlicher Interpretation, Erfahrung und Empathie geprägt?
- Wie hoch sind die Risiken bei Fehlentscheidungen, und wen würden diese betreffen?
- Welche Teile des Prozesses erfordern zwingend menschliches Urteilsvermögen?

Je klarer ein Prozess strukturiert, wiederholbar und regelbasiert ist, desto eher eignet er sich für eine Form der Automatisierung. Je stärker dagegen Intuition, Beziehungsarbeit oder situatives Fingerspitzengefühl eine Rolle spielen, desto vorsichtiger sollte eine Automatisierung geprüft werden.

Wie stark sollte ein Task automatisiert werden?

Selbst wenn ein Prozess grundsätzlich für Automatisierung geeignet erscheint, bleibt eine zentrale Frage offen: **Wie stark sollte er automatisiert werden?**

Hierfür bietet das Konzept der **Levels of Automation (LoA)** von Parasuraman et al. eine theoretische Grundlage. Es beschreibt, wie sich unterschiedlich stark automatisierte Systeme auf die Rolle, Wahrnehmung und Leistungsfähigkeit von Menschen auswirken.

Automatisierung		Die Maschine (der Computer)...
Hoch	10	entscheidet alles, handelt autonom, ignoriert den Menschen
	9	informiert den Menschen nur dann, wenn sie es für richtig hält
	8	informiert den Menschen nur dann, wenn er fragt
	7	handelt automatisch, informiert den Menschen danach
	6	gibt dem Menschen eingeschränkt Zeit, einen automatischen Prozess zu stoppen
	5	führt eine vorgeschlagene Handlung aus, wenn der Mensch zustimmt
	4	schlägt eine Handlung vor
	3	stellt ein vorselektiertes Set an Entscheidungs-/Handlungsalternativen bereit
	2	stellt ein vollständiges Set an Entscheidungs-/Handlungsalternativen bereit
Gering	1	hilft nicht - der Mensch muss alles selbst entscheiden und ausführen

Ein klassisches Beispiel ist der Unterschied zwischen:

- einem Assistenzsystem, das lediglich warnt (z. B. beim Verlassen der Fahrspur), und
- einem System, das selbst aktiv in die Lenkung eines Fahrzeugs eingreift.

In der Forschung werden häufig **zehn Abstufungen von sehr geringer bis zu vollständiger Automatisierung** unterschieden. Diese Abstufungen lassen sich auf vier Hauptbereiche menschlicher Informationsverarbeitung übertragen.

Vier Bereiche der Automatisierung entlang der menschlichen Informationsverarbeitung

Information Acquisition - Informationsbeschaffung

Diese Ebene bezieht sich auf das **Sammeln und Registrieren von Eingangsdaten**. Automatisierung in diesem Bereich unterstützt menschliche sensorische Prozesse, indem Daten schneller, vollständiger oder kontinuierlicher erfasst werden.

Beispiel im Organisationskontext: Ein System sammelt automatisch Anmeldungen, Standortdaten oder Verfügbarkeiten von Freiwilligen, anstatt dass diese manuell zusammengetragen werden.

Information Analysis - Informationsanalyse

Hier geht es um **kognitive Funktionen** wie Vergleichen, Strukturieren, Extrapolieren oder Vorhersagen. Systeme können Muster erkennen, Zusammenhänge herstellen oder Daten vorverarbeiten.

Beispiel: Ein System analysiert eingehende Daten und schlägt auf dieser Basis mögliche Einsatzstellen für Freiwillige vor.

Decision & Action Selection - Entscheidungs- und Handlungsauswahl

In dieser Phase werden **Handlungsoptionen abgewogen und ausgewählt**. Eine Automatisierung bedeutet hier, dass das System bereits Entscheidungen trifft oder stark vorgibt - mit oder ohne menschliche Bestätigung.

Beispiel: Das System entscheidet selbstständig, welcher Freiwillige welcher Einsatzstelle zugewiesen wird, anstatt nur Vorschläge zu liefern.

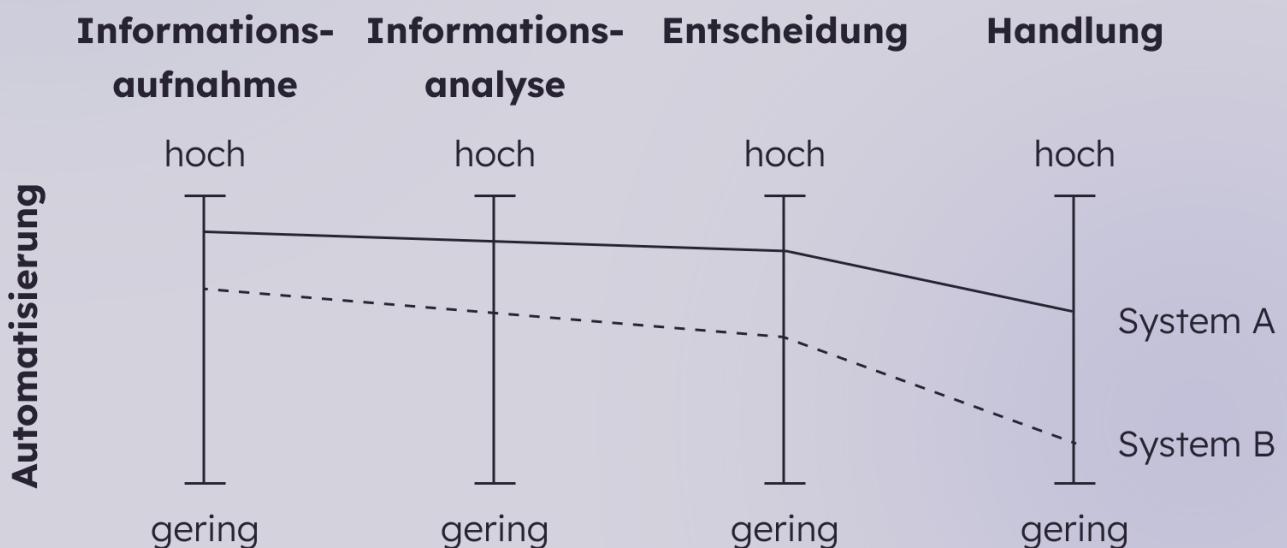
Action Implementation - Handlungsausführung

Diese letzte Phase betrifft die **eigentliche Umsetzung der Entscheidung**. Automatisierung ersetzt hier in der Regel direkt die physische oder kommunikative Handlung eines Menschen.

Beispiel: Ein System versendet automatisch Zusagen, Termine und Zugangsdaten an Freiwillige und Einsatzstellen.

Wichtig ist: Ein System hat nicht ein einziges Automatisierungslevel, sondern kann in den einzelnen Phasen unterschiedlich stark automatisiert sein.

So kann beispielsweise die Datenerfassung vollständig automatisiert sein, während die Entscheidung bewusst beim Menschen verbleibt. Moderne Systeme können zudem so gestaltet werden, dass sie ihre Automatisierungsstufe situativ anpassen.



Für die Systemgestaltung bedeutet das:

- Für jeden Teilschritt eines Prozesses sollte die **passende Automatisierungsstufe** identifiziert werden.
- Anschließend sollte reflektiert werden: **Welche Auswirkungen hat diese Entscheidung auf die beteiligten Menschen?**

Dabei sind insbesondere vier Aspekte relevant

1. Mental Workload (mentale Arbeitsbelastung)

Höhere Automatisierungsstufen können die Belastung von Nutzer:innen reduzieren, etwa indem große Datenmengen automatisch ausgewertet und visualisiert werden. Aufgaben müssen dann nicht mehr manuell erledigt werden.

Gleichzeitig gilt: Eine höhere Automatisierung senkt die mentale Belastung nicht automatisch. Sie kann diese je nach Gestaltung des Systems sogar erhöhen - etwa dann, wenn Informationen unübersichtlich präsentiert werden oder Entscheidungen nicht mehr nachvollziehbar sind.

2. Situation Awareness (Situationsbewusstsein)

Das Situationsbewusstsein beschreibt, wie gut Menschen Veränderungen und Zustände eines Systems wahrnehmen und verstehen.

In hochautomatisierten Systemen sinkt dieses Bewusstsein häufig, weil Menschen nur noch eine **überwachende Rolle** einnehmen. Veränderungen werden schlechter erkannt, insbesondere wenn sich ein System „unauffällig“ verhält.

Für sicherheitskritische oder sensible Bereiche - wie sie in gemeinwohlorientierten Organisationen häufig vorkommen - ist dieses Risiko besonders relevant.

3. Complacency (Übervertrauen)

Wenn ein System in der Regel zuverlässig arbeitet, neigen Menschen dazu, es nicht mehr kritisch zu hinterfragen. Fehler oder Ausnahmen werden dann häufig übersehen.

Es entsteht ein **Übervertrauen in die Technik**, das nicht im Verhältnis zu ihren tatsächlichen Fähigkeiten steht. Gerade bei KI-Systemen, deren Entscheidungsgrundlagen oft schwer nachvollziehbar sind, kann dies zu problematischen Abhängigkeiten führen.

4. Skill Degradation (Abbau menschlicher Fähigkeiten)

Werden Aufgaben über längere Zeit von einem System übernommen, verlieren Menschen nach und nach die Fähigkeit, diese selbstständig auszuführen. Das wird besonders problematisch, wenn sie im Notfall plötzlich wieder eingreifen oder einspringen müssen.

In Organisationen, die stark auf Wissen, Erfahrung und Beziehungskompetenz angewiesen sind, kann dies zu einem langfristigen Kompetenzverlust führen.

Preferred Automation Tasks Scale (PATS)

Kursübersicht > Automatisierungspotenziale erkennen

Neben technischen und organisatorischen Überlegungen ist es entscheidend zu verstehen, **was die Menschen in Ihrer Organisation überhaupt möchten.**

Die **PATS-Skala (Preferred Automation Task Scale)** wurde entwickelt, um zu messen, welche Aufgaben Menschen lieber von einem automatisierten System und welche sie lieber von anderen Menschen durchführen lassen möchten. Ein solches Tool gibt die Möglichkeit innerhalb einer Organisation zu überprüfen wie die verschiedenen Stakeholder eine Automatisierung des Prozesses bewerten würden. Sie basiert auf den vier zentralen Funktionen von Automatisierung nach Parasuraman et al.:

- 1 **Daten sammeln** - Informationen erfassen, filtern und organisieren
- 2 **Daten analysieren** - Muster erkennen und Schlussfolgerungen ziehen

3 Entscheidungen treffen - Optionen abwägen und auswählen

4 Handlungen umsetzen - Entscheidungen praktisch durchführen

Die Skala besteht aus **zwölf Fragen**, die auf einer **sechsstufigen Bewertungsskala** beantwortet werden:

- von „bevorzuge vollständig menschliche Ausführung“
- bis „bevorzuge vollständig automatisierte Ausführung“

		Bevorzuge voll und ganz Menschen	Bevorzuge überwiegend Menschen	Bevorzuge eher Menschen	Bevorzuge automatisierte Systeme	Bevorzuge überwiegend automatisierte Systeme	Bevorzuge voll und ganz automatisierte Systeme
Acq2	Daten anhand von Kriterien organisieren.						
Acq3	Daten, die wichtig sein könnten, hervorheben.						
Acq4	Daten, die möglicherweise nicht relevant sind, herausfiltern.						
Ana2	Daten verarbeiten, um neue Informationen zu erhalten.						
Ana3	Lücken in den Daten mit Hilfe verfügbarer Informationen füllen.						
Ana4	Daten aus verschiedenen Quellen kombinieren, um Schlussfolgerungen zu ziehen.						
Dec2	Eine Entscheidung darüber treffen, was zu tun ist, nach einem Vergleich der Kosten und Nutzen verschiedener Optionen.						
Dec3	Mit Hilfe von Schlussfolgerungen eine Entscheidung darüber, was zu tun ist, treffen.						
Dec4	Eine Entscheidung darüber treffen, was zu tun ist, je nachdem, welche Bedingungen erfüllt sind.						
Act1	Die beschlossene Handlung durchführen.						
Act2	Werkzeuge oder Hardware benutzen, um die gewählten Handlungen auszuführen.						
Act3	Verschiedene Aktionen kombinieren, um die gewählten Handlungen auszuführen.						

Die Ergebnisse können in verschiedenen Szenarien verglichen werden (z. B. „normaler Betrieb“ vs. „akuter Personalmangel“), um für die beteiligten Personen durchzuspielen, wie sich unterschiedliche Bedingungen auf ihre Antworten auswirken. Diese szenarienbasierte Strategie findet häufig in sicherheitskritischen Bereichen wie Medizin oder Luftfahrt Anwendung, ist aber auch für gemeinwohlorientierte Organisationen sehr wertvoll.

Wo kann der Einsatz der PATS helfen?

Der Einsatz von PATS kann dabei helfen:

- Systeme bedürfnisorientiert zu gestalten
- Transparenz und Beteiligung zu fördern
- frühzeitig Konflikte oder Bedenken sichtbar zu machen
- Potentiale für Automatisierung zu priorisieren

Wenn Sie die PATS selbst einsetzen möchten, um eine Umfrage in Ihrer Organisation durchzuführen, finden Sie hier eine fertige Druckversion:

[PATS - Deutsche Druckversion](#)

Entscheidung über den Einsatz von **KI: Wann ist KI sinnvoll?**

Kursübersicht > Automatisierungspotenziale erkennen

Nachdem ein Potenzial für Automatisierung identifiziert wurde, stellt sich eine weiterführende, kritischere Frage: **Ist für diese Aufgabe wirklich Künstliche Intelligenz notwendig oder reicht eine klassische, regelbasierte Automatisierung aus?**

Dabei gilt die Unterscheidung:

- **Automatisierung:** Klare, vordefinierte Regeln und Abläufe („Wenn X, dann Y“)
- **KI:** Lernen aus Daten, Erkennung von Mustern und Zusammenhängen, Umgang mit Unsicherheit und Variabilität, aber eben auch höhere rechtliche Anforderungen, geringere Transparenz und höhere Kosten.

Um diese Frage fundiert zu beantworten, können Sie folgende Checkliste nutzen:

Checkliste: Eignung von KI-Technologien

1

Gibt es ausreichend digitale Daten? Sind in Menge und Qualität genügend Daten vorhanden, aus denen ein System lernen kann?

2

Steht Mustererkennung im Fokus (statt starrer Regeln)? Geht es darum, Zusammenhänge, Tendenzen oder Abweichungen zu identifizieren, anstatt feste Entscheidungsbäume abzuarbeiten?

3

Kann das Ergebnis überprüft oder erklärt werden? Ist Nachvollziehbarkeit möglich (z. B. durch Vergleich, Plausibilitätsprüfung, menschliche Kontrolle)?

4

Sind Datenschutz und Bias handhabbar? Können Risiken im Umgang mit personenbezogenen Daten und Verzerrungen (Bias) realistisch kontrolliert werden?

5

Ist das System langfristig wartbar und verantwortbar? Gibt es Ressourcen, Wissen und Zuständigkeiten für Pflege, Monitoring und Weiterentwicklung?

6

Entspricht ein mögliches System den rechtlichen Rahmenbedingungen? (z. B. DSGVO, AI Act, organisationsinterne Richtlinien)

Neben der technischen Machbarkeit ist die menschliche Perspektive entscheidend. Die folgenden Fragen helfen, den Nutzen und die Risiken aus Sicht der Nutzenden zu reflektieren:

1. Nützlichkeit

- Welche konkreten Ergebnisse könnte ein KI-System erzeugen?
- Auf welche Weise könnte es diesen Prozess erleichtern, beschleunigen oder verbessern?
- Welches Problem würde ohne KI bestehen bleiben?

2. Rahmenbedingungen der Zusammenarbeit

- Wie wird die Zusammenarbeit mit dem KI-System praktisch aussehen?
- Welche Aufgaben übernimmt das System - welche verbleiben beim Menschen?
- Auf welche Informationen haben Mensch und System Zugriff?
- Wo bestehen Informationsunterschiede oder -grenzen?

3. Zufriedenheit und Risiken

- Wo könnten in der geplanten Kooperation Pain Points entstehen?
- Welche Risiken bestehen im Umgang mit dem System?
- Welche Risiken ergeben sich für die Zusammenarbeit im Team?
- Wo könnten Unsicherheiten, Misstrauen oder Überforderung entstehen?

Fazit

Ob KI sinnvoll ist, hängt davon ab, ob genügend qualitativ gute Daten vorliegen, Mustererkennung statt fester Regeln erforderlich ist, Ergebnisse überprüfbar bleiben und Datenschutz, Bias sowie rechtliche Vorgaben handhabbar sind. Außerdem muss das System langfristig wartbar sein.

Neben der Technik zählt die Perspektive der Nutzenden: Bringt KI einen klaren Nutzen, passt sie gut in die Arbeitsabläufe, und entstehen weder neue Risiken noch Belastungen für Menschen oder Teams?

05

Quellen

Kursübersicht > Automatisierungspotenziale erkennen

Literaturverzeichnis

- Bieber, C., & für politische Bildung), *bpb* (Bundeszentrale. (n.d.).
Digitalisierung. <https://www.bpb.de/kurz-knapp/lexika/handwoerterbuch-politisches-system/511460/digitalisierung/>
- Bundesamt für Sicherheit in der Informationstechnik (BSI). (n.d.).
Künstliche Intelligenz – wir bringen Ihnen die Technologie näher.
https://www.bsi.bund.de/DE/Themen/Verbraucherinnen-und-Verbraucher/Informationen-und-Empfehlungen/Technologien_sicher_gestalten/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html
- Bundeszentrale für politische Bildung (bpb). (n.d.). *Automatisierung*.
<https://www.bpb.de/kurz-knapp/lexika/lexikon-der-wirtschaft/18743/automatisierung/>
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2), 230–253.
<https://doi.org/10.1518/001872097778543886>

- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3), 286-297. <https://doi.org/10.1109/3468.844354>
- Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (Fourth edition). Pearson.
- Zoubir, M., Gruner, M., Schrills, T., & Franke, T. (2024). *Development and evaluation of the preference for automation types scale*. <https://doi.org/10.31219/osf.io/fwa8m>

Modul:

06 EU AI Act

Gefördert vom:



Bundesministerium
für Bildung, Familie, Senioren,
Frauen und Jugend



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MULTIMEDIALE
UND INTERAKTIVE SYSTEME

01

Einleitung

[**Kursübersicht**](#) > [**EU AI Act**](#)

In diesem Modul haben wir den EU AI Act für Sie aufbereitet – das ist die europäische Verordnung über künstliche Intelligenz, laut derer KI-Systeme künftig vier Risikogruppen zugewiesen werden sollen. Das geschieht in Abhängigkeit vom Nutzungskontext des Systems und kann schwerwiegende Folgen für Unternehmen haben, die KI-Systeme entwickeln. Hier erfahren Sie, was sich durch das Gesetz verändert wird und welche Auswirkungen das neue Gesetz auf Entwickelnde haben kann. Abschließend diskutieren unsere Experten die Chancen und Herausforderungen, die der EU AI Act für Developer und Nutzende von KI-Systemen darstellt.

1. Was ist der EU AI Act?

Nachdem Sie nun ein kurzes Einführungsvideo gesehen haben, wenden wir uns nun einer etwas detaillierteren Betrachtung des EU AI Acts zu.

Ziele des EU AI Acts

Mit dem EU AI Act haben sich die Mitgliedstaaten der EU zum Ziel gesetzt, das erste umfassende Regularium für den Umgang mit Systemen, die auf Künstlicher Intelligenz (KI) basieren, zu verfassen. Ziel ist es, rechtliche Rahmenbedingungen zu schaffen, in denen KI-Projekte in Wirtschaft, Forschung und Gesellschaft so eingesetzt und entwickelt werden können, dass ein vertrauenswürdiger Umgang mit den Systemen möglich ist. Die fundamentalen Rechte der Nutzer:innen, aber auch Aspekte wie Sicherheit und das Handeln auf Grundlage ethischer Prinzipien, sollen dabei mit einbezogen werden. Dabei soll der EU AI Act über die Grenzen der EU hinaus einen Impuls setzen und, ähnlich wie die DSGVO im Datenschutz, einen Standard stellen, der auch in Nicht-EU-Ländern wie den USA oder Japan genutzt wird.

Entwicklung des Acts

Bei der Entwicklung des EU AI Acts handelt es sich um einen komplexen bürokratischen Prozess, der sich nicht nur über viele Jahre hinweg entwickelte, sondern auch große Teile des parlamentarischen EU-Apparats durchlaufen hat.

Eine vollständige Übersicht des zeitlichen Verlaufs des Gestaltungsprozesses und der damit verbundenen Institutionen und Zwischenstände finden Sie hier:

<https://artificialintelligenceact.eu/de/entwicklungen/>.

2. Vergleich mit anderen Ländern / Regionen

insbesondere England, USA und Kanada

Betrachtet man den EU AI Act als groß angelegte Normierung innerhalb des europäischen Raums, drängt sich schnell die Frage auf, welche Auswirkungen die Gesetzgebung außerhalb seiner Mitgliedstaaten haben wird. Es kann dabei davon ausgegangen werden, dass das EU-Parlament zwar primär die Regulierung im eigenen Legislaturbereich im Blick hat, aber auch auf andere große Volkswirtschaften wie die USA, China und das Vereinigte Königreich schaut, wenn es um den Einsatz von KI geht. Wie schon bei der Datenschutz-Grundverordnung (DSGVO) scheint hier der Gedanke zu sein, einen weitreichenden "Goldstandard" zu schaffen, der auch die Gesetzgebung in den Nationen außerhalb der EU bestimmt. Wir halten es deshalb für sinnvoll, einen kurzen Blick auf den aktuellen Stand der KI-Regulierung in anderen Nationen zu werfen, um diese Vorstellung besser einordnen zu können.

Vereinigtes Königreich

Schaut man dazu beispielsweise über den Kanal ins Vereinigte Königreich, so stellt man fest, dass auch dort weitreichende Maßnahmen für die Regulierung und den ethischen Umgang mit KI bereits getroffen worden sind. Die britische Regierung setzt dabei auf bestehende sektorale Vorgaben wie bspw. die KI-Prinzipien der OECD oder die Empfehlung zum ethischen Umgang mit KI der UNESCO. Diese übergeordneten Richtlinien werden durch lokal angetriebene Maßnahmen erweitert. Von besonderer Bedeutung ist dabei zum Beispiel die Bletchley Declaration aus November 2023, bei der sich 28 Länder, darunter die Vereinigten Staaten, China und die Europäische Union, geeinigt haben, international bei der Bewältigung von Herausforderungen und Risiken im Bereich der KI zusammenzuarbeiten. Im Fokus standen dabei vor allem "frontier"-Systeme, also KI-Grundlagenmodelle, die für alle möglichen Anwendungsfälle nutzbar gemacht werden können, so wie bspw. die ChatGPT zugrunde liegenden LLMs. Es gibt also ein klares Bewusstsein für die Bedeutung des Themas KI und erste Bestrebungen für Lösungen. Die dabei getroffenen Vereinbarungen sind dabei eher Leitlinien und weniger strenges Regularium, als es der EU AI Act sein möchte.

USA

Demgegenüber steht die Regulierung von KI in den USA derzeit noch am Anfang. Zwar sind auch diese Teil der Bletchley Declaration, trotzdem fehlt ein kohärentes nationales Regelwerk. Aktuell gibt es in den USA keine umfassende föderale KI-Regulierung, sondern lediglich fragmentierte Richtlinien und Bestrebungen auf Bundesstaatenebene und in verschiedenen Sektoren. Bundesbehörden wie die Federal Trade Commission (FTC) haben zwar Leitlinien zur Vermeidung unfairer oder

irreführender KI-Anwendungen herausgegeben, aber umfassende gesetzliche Regelungen stehen noch aus.

Allerdings gibt es zunehmend Bestrebungen, eine konsistente Regulierung zu entwickeln. Präsident Joe Biden hat Anfang 2021 den "National Artificial Intelligence Initiative Act" unterzeichnet, der die Forschung und Entwicklung von KI koordiniert und fördert. Zudem hat das Weiße Haus Ende 2022 einen "Blueprint for an AI Bill of Rights" veröffentlicht, der Prinzipien zum Schutz der Bürgerrechte im Zusammenhang mit KI vorschlägt. Während die EU mit dem AI Act einen klaren und strengen Regulierungsrahmen vorgibt, arbeiten die USA daran, ihre Strategie zu entwickeln, die wahrscheinlich stärker auf Selbstregulierung und sektorale Ansätze setzt, um Innovationen nicht zu behindern (Pinsent Masons) (Skadden, Arps, Slate, Meagher & Flom LLP).

Kanada

Im Kontext des EU AI Acts hat Kanada ebenfalls Schritte unternommen, um den Einsatz von KI zu regulieren und zu fördern. Aktuell wird KI in Kanada hauptsächlich durch den Artificial Intelligence and Data Act (AIDA) reguliert, der Teil des umfassenden Digital Charter Implementation Acts ist, welcher im Juni 2022 vorgeschlagen wurde. AIDA zielt darauf ab, KI-Systeme zu regulieren, die ein erhebliches Risiko für die Sicherheit der Menschen oder ihre Grundrechte darstellen. Der Ansatz umfasst Verpflichtungen zur Transparenz, zur Risikobewertung und zur Einhaltung ethischer Standards.

Zusätzlich gibt es in Kanada Bestrebungen, die Regulierung weiter zu verfeinern und zu stärken. Die kanadische Regierung arbeitet daran, Richtlinien und Standards zu entwickeln, die sicherstellen, dass KI-

Systeme sicher, fair und transparent sind. Dies beinhaltet auch die Zusammenarbeit mit internationalen Partnern und Organisationen, um globale Standards zu fördern und die Interoperabilität von Regulierungsrahmen zu gewährleisten. Im Vergleich zum EU AI Act, der einen sehr strukturierten und strengen Rahmen vorgibt, verfolgt Kanada einen eher kooperativen und flexiblen Ansatz. Die kanadische Regulierung konzentriert sich auf die Förderung von Innovationen, während sie gleichzeitig sicherstellt, dass die Entwicklung und der Einsatz von KI ethisch und verantwortungsvoll erfolgen (Pinsent Masons) (Skadden, Arps, Slate, Meagher & Flom LLP).

Zusammengefasst kann festgehalten werden, dass zwar viele Nicht-EU-Nationen eine klare Vorstellung von KI und potenziellen unkontrollierten Auswirkungen haben, im Gegensatz zur EU allerdings teilweise noch am Anfang einer konkreten Ausformulierung von Regeln und Gesetzen stehen oder generell einen offeneren Ansatz mit Blick auf möglichst freie Innovationsentwicklung verfolgen.

3. Umsetzung des Acts in den Mitgliedstaaten

Die faktische Umsetzung des EU AI Acts in den Mitgliedstaaten erfordert sorgfältige Planung und Koordination, um die umfassenden Anforderungen des Gesetzes zu erfüllen.

Primäre Inhalte des EU AI Acts

Der EU AI Act ist ein umfassendes Regelwerk, das den Einsatz von KI innerhalb der EU reguliert. Zu den primären Inhalten gehören die Kategorisierung von KI-Systemen nach ihrem Risiko (unzulässiges, hohes, begrenztes und minimales Risiko), spezifische Anforderungen für Hochrisiko-Systeme, Transparenz- und Sicherheitsanforderungen sowie die Einrichtung eines EU-weiten Überwachungssystems für KI-Anwendungen. Einen tieferen Einblick in das Thema Risikoklassifizierung und Risikostufen finden Sie in den Kapiteln 02 und 03. Diese Maßnahmen zielen darauf ab, die Sicherheit, Transparenz und Verantwortung im Umgang mit KI zu gewährleisten und gleichzeitig Innovationen zu fördern. Darüber hinaus bilden sie eine feste rechtliche Grundlage, die Unternehmen, Forschung und Endnutzer:innen befähigt, rechtssicher mit KI umzugehen.

Anwendungsbereich und betroffene Systeme

Der Act findet Anwendung bei verschiedenen KI-Systemen, abhängig von ihrem Risiko. Hochrisiko-Systeme umfassen beispielsweise KI-Anwendungen in kritischen Infrastrukturen, wie etwa KI-gesteuerte Systeme im Gesundheitswesen. Ein konkretes Beispiel wäre ein KI-gestütztes Diagnosetool in Krankenhäusern, das strengen Auflagen hinsichtlich Datenqualität, Transparenz und menschlicher Aufsicht unterliegt. Niedrigrisiko-Systeme, wie etwa Chatbots oder KI-basierte Spiele, unterliegen weniger strengen Regelungen, müssen aber dennoch gewisse Transparenzanforderungen erfüllen.

Relevante Beteiligte

Alle Anbieter:innen und Nutzer:innen von KI-Systemen innerhalb der EU müssen sich mit den Anforderungen des Acts auseinandersetzen. Dies umfasst Entwickler:innen, Anbieter:innen und Anwender:innen von KI-Technologien. Privatpersonen, die KI-Anwendungen wie ChatGPT nutzen, sind in der Regel nicht direkt betroffen, solange sie diese nur als Endnutzer:innen einsetzen und die Anwendungen den regulatorischen Anforderungen entsprechen. Unternehmen, die solche Technologien entwickeln oder bereitstellen, müssen hingegen sicherstellen, dass ihre Produkte konform sind.

Sanktionen bei Verstößen

Bei Verstößen gegen den EU AI Act drohen erhebliche Sanktionen. Die vorgeschlagenen Strafen umfassen Bußgelder von bis zu 35 Millionen Euro oder 7% des weltweiten Jahresumsatzes. Die faktische Umsetzung des EU AI Acts in den Mitgliedstaaten erfordert umfassende Maßnahmen zur Einhaltung der neuen Vorschriften, die eine sichere und transparente Nutzung von KI sicherstellen sollen.

4. Fazit

Zusammenfassend lässt sich sagen, dass der EU AI Act ein umfassendes und wegweisendes Regelwerk darstellt, das den Einsatz von KI in der EU regulieren soll. Er kategorisiert KI-Systeme nach ihrem Risikoniveau und legt spezifische Anforderungen für Hochrisiko-Systeme fest, um Sicherheit, Transparenz und Verantwortlichkeit zu gewährleisten. Der Act betrifft eine Vielzahl von KI-Anwendungen, von Gesundheitsdiagnosetools bis hin zu Chatbots, und erfordert von Anbieter:innen und Nutzer:innen die Einhaltung strenger Vorgaben. Bei Verstößen drohen erhebliche Sanktionen, einschließlich hoher Bußgelder. In den nächsten Kapiteln werden wir uns detailliert mit den Risikostufen und der Klassifizierung von KI-Systemen befassen, um ein tieferes Verständnis für die Implementierung und Einhaltung des EU AI Acts zu entwickeln.

Risikostufen - Anwendungsbeispiele

Kursübersicht > [EU AI Act](#)

Hier werden mögliche Anwendungen beispielhaft betrachtet und welche Auswirkungen der EU AI Act auf diese hat.

1. Einteilung in Risikostufen anhand Anwendungsbeispielen

Nachdem wir zuvor einen groben Überblick über die Entwicklung und die wichtigsten Aspekte des EU AI Acts gegeben haben, möchten wir nun einen genaueren Blick auf das Herzstück des EU AI Acts werfen: die Einteilung der Risikostufen. In diesem Abschnitt wird erläutert, welche Risikostufen es gibt, welche Systeme in welche Kategorie fallen und welche Anforderungen daraus für Organisationen entstehen. Sobald Sie ein grundlegendes Verständnis der Risikostufen erlangt haben, werden wir im zweiten Schritt zwei praktische Beispiele für die Einordnung von Systemen in die verschiedenen Risikostufen betrachten und Ihnen ein Tool vorstellen, das Sie selbst zur Einstufung nutzen können.

2. Die Beispiele

Nachdem Sie sich nun mit den Grundlagen der Risikostufen vertraut machen konnten, können wir uns nun den Beispielen zuwenden, die praktischen Herausforderungen besser illustrieren.

Beispiel 1 - Antrags Assistent

Stellen Sie sich vor, Sie sind Teil einer kleinen in Berlin ansässigen Organisation, die es sich zum Ziel gesetzt hat Personen aus benachteiligten Gruppen bei der Kommunikation mit Behörden zu unterstützen, bspw. durch Hilfe beim Schreiben von Briefen oder Anträgen, kleiner Übersetzungsleistungen o.Ä.. Um Ihre Prozesse zu optimieren haben Sie vor ein Unterstützungstool einzukaufen, dass die Unterlagenprüfung für Sie übernimmt. Personen, die zu Ihnen kommen können dort ihre Dokumente digital hinterlegen, diese werden dann vom Antrags Assistenten geprüft, der Ihnen und ihren Kolleg:innen eine Auskunft darüber gibt, wie das System die Chancen auf Erfolg bei Antragsstellung bewertet. Das System kann keine Personen ablehnen und keine eigenständigen Entscheidungen treffen.

Was glauben Sie? Nehmen Sie sich einen Moment Zeit und denken Sie darüber nach wo ein solches System eingeordnet werden könnte. Wir fassen die wichtigsten Informationen hier noch mal zusammen, dabei spielen nicht nur die Dinge eine Rolle, die das System tut, sondern explizit auch was es nicht tut oder kann.

Übersicht Beispielsystem 1 - Antrags Assistent

- Unsere Organisation setzt das System nur ein.
- Wir sind mit unserem Standort in Berlin innerhalb der EU niedergelassen.
- Wir nutzen das System weder für militärische Zwecke, noch sind wir Teil einer Behörde oder Forschungseinrichtung.
- Das System wird nicht für Dinge wie Social Scoring, Emotionserkennung oder Verhaltensmanipulation genutzt.
- Da es sich um ein System handelt, dass potenziell den Zugang zu privaten und öffentlichen Leistungen beeinflusst, könnte dies besondere Auswirkungen auf die Einstufung unseres Tools haben. Wichtig ist dabei vor allem, dass kein erhebliches Risiko für die Gesundheit, die Sicherheit oder die Grundrechte einer Person darstellt.

Prüfen wir basierend auf diesen Informationen unser Assistenzsystem so bedeutet das, dass die Anwendung mit hoher Wahrscheinlichkeit der niedrigsten Risikostufe zugeordnet wird. Für Sie als Nutzende heißt das, dass Sie das System wie geplant nutzen können. Auf Anbieterseite ist dies allerdings mit einigen Pflichten verbunden. So muss das System in einer EU-Datenbank registriert werden und die getätigten Anfragen müssen beim Anbieter so gesichert werden, dass dieser Sie auf Anfrage der EU-Behörden übertragen kann.

Beispiel 2

Für unser zweites Beispiel stellen wir uns vor wir sind Teil einer in den USA und Europa agierenden auf Nachhaltigkeit und Tierwohl ausgerichtete Organisation, die in Kooperation mit einer ländlichen Kommune die Population in einem Waldstück überwachen und messen möchte. Um ein möglichst detailliertes Bild zu bekommen, sendet Ihnen ihre Hauptstelle in den Staaten durch KI-Erkennungssystem gesteuerte Kameras, die im Waldstück angebracht werden und automatisch Tiere bedrohter Arten identifizieren und die gemachten Bilder speichern sollen. Bei der genutzten Tiererkennungssoftware handelt es sich um eine etablierte Anwendung eines großen Technologiehauses, die in Kooperation mit ihrer Hauptstelle entwickelt und in den USA schon weitläufig mit Erfolg eingesetzt wird. Aufgrund der Serverinfrastruktur, auf die das System zurückgreift werden Daten, die in Deutschland aufgenommen werden in der Hauptstelle in den USA verarbeitet und aufgenommen.

Nehmen Sie sich auch hier wieder einen Moment Zeit und denken Sie darüber nach welcher Risikostufe so ein System zugeordnet werden könnte. Hier noch einmal die wichtigsten Informationen:

Übersicht Beispielsystem 2

- Das System wird von unserer Organisation hergestellt und eingesetzt.
- Wir sind eine amerikanische Organisation mit einem Standort in Deutschland.
- Wir nutzen das System weder für militärische Zwecke, noch sind wir Teil einer Behörde oder Forschungseinrichtung.
- Das System wird nicht für Dinge wie Social Scoring, Emotionserkennung oder Verhaltensmanipulation genutzt.
- Das System wird nicht in einem Hochrisikobereich eingesetzt bspw. bei der Strafverfolgung oder zum Grenzkontrollmanagement.

Prüfen wir basierend auf diesen Informationen unser Assistenzsystem so bedeutet das, dass auch diese Anwendung mit hoher Wahrscheinlichkeit der niedrigsten Risikostufe zugeordnet wird. Trotz der Verarbeitung der Daten im Nicht-Eu-Ausland und dem Einsatzgebiet bei der Überwachung handelt es sich mit hoher Wahrscheinlichkeit um ein System, dass auf Nutzendenseite keine weiteren Pflichten mit sich bringt. Unsere Dachorganisation bzw. der Technologiekonzern unterliegen dabei trotz ihres Standortes in den USA den gleichen Transparenzpflichten wie in Beispiel 1, wenn sie ihr System innerhalb der EU einsetzen wollen.

Die genaue Einschätzung in welche Risikostufe ein gegebenes KI-System fällt, kann durchaus komplex sein und hängt von mehr Faktoren ab als wofür es inhaltlich genutzt wird bspw. ob ich als Hersteller und Nutzender auftrete. Wo ich das System einsetzen will? Ob ich Teil der EU bin etc.

3. EU AI Act Compliance Checker

Um sich einen Überblick über die verschiedenen Möglichkeiten zu verschaffen, gibt es den EU-Compliance-Checker. Das Tool bietet die Möglichkeit verschiedene Varianten durchzuspielen, um herauszufinden welche Regelungen für das eigene System gelten. Wir empfehlen daher es einmal selbst auszuprobieren und die oben genannten Beispiele oder eigene Idee einfach mal auf <https://artificialintelligenceact.eu/de/bewertung/eu-ai-act-compliance-checker/> oder mit einem Klick auf die unten befindliche Einbindung zu testen.

EU AI Act Compliance Checker

Das EU-Gesetz über künstliche Intelligenz führt neue Verpflichtungen für Unternehmen innerhalb und außerhalb der EU ein. Nutzen Sie unser interaktives Tool, um festzustellen, ob Ihr KI-System davon betroffen ist oder nicht.

Wenn Sie über Ihre Verpflichtungen im Rahmen des EU AI Act auf dem Laufenden bleiben möchten, empfehlen wir Ihnen, den EU AI Act Newsletter zu abonnieren.

Um mehr Klarheit zu schaffen, empfehlen wir Ihnen, sich rechtlich beraten zu lassen und die nationalen Richtlinien zu befolgen. Weitere Informationen über die Durchsetzung des EU-AI-Gesetzes in Ihrem Land werden voraussichtlich im Jahr 2024 zur Verfügung gestellt.

Feedback - Wir arbeiten an der Verbesserung dieses Tools. Bitte senden Sie Ihr Feedback an Taylor Jones unter taylor@futureoffice.org

Sehen Sie sich den offiziellen Text an, oder durchsuchen Sie ihn online mit unserem AI Act Explorer. Der in diesem Tool verwendete Text ist das "Gesetz über künstliche Intelligenz (Verordnung (EU) 2024/1689), Fassung des Amtsblatts vom 13. Juni 2024". Interinstitutionelle Akte: 2021/0106(COD)

Wie wirkt sich das EU-KI-Gesetz auf mein KI-System aus?

Bitte füllen Sie dieses Formular für jedes einzelne in Ihrer Organisation verwendete AI-System aus.

Ist mein System ein "KI-System" im Sinne des EU-KI-Gesetzes?

Ein System der künstlichen Intelligenz (KI-System) ist definiert als: Ein maschinengestütztes System, das so konzipiert ist, dass es mit unterschiedlichem Grad an Autonomie operieren kann und nach seiner Einführung Anpassungsfähigkeit zeigt, und das für explizite oder implizite Ziele aus den Eingaben, die es erhält, ableitet, wie es Ergebnisse wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erzeugen kann, die physische oder virtuelle Umgebungen beeinflussen können.

Quelle: Artikel 3, Punkt 1

Entitätstyp

Welche Art von Einrichtung ist Ihre Organisation?

Hinweis: Wenn Sie die Definition mehrerer Entitätstypen erfüllen, müssen Sie das Formular mehrfach ausfüllen, einmal für jeden Entitätstyp.

- Anbieter
- Einsetzer
- Verteiler
- Importeur
- Hersteller des Produkts
- Bevollmächtigter

[Siehe Definitionen](#)

Definitionen für diese Begriffe anzeigen.

Quelle: Artikel 3 Nummern 2-8, Erwägungsgrund 87

Im nächsten Abschnitt gehen wir dann noch mal konkret auf die Auswirkungen der Risikostufen auf mögliche Entwicklungsprozesse ein.

Risikostufen - Auswirkungen

Kursübersicht > [EU AI Act](#)

In diesem Kapitel werden einzelne Artikel des Acts näher betrachtet und welche Auswirkungen sie auf verschiedene System haben können.

1. Was muss je nach Risikostufe beachtet werden?

Dieser Text zielt darauf ab, ein tieferes Verständnis der verschiedenen Risikokategorien zu vermitteln, die in der KI-Verordnung (KIVO) definiert sind, die Anforderungen für jede Kategorie zu erläutern, mit besonderem Fokus auf Hochrisiko-KI-Systeme, und das Konzept eines KI-Managementsystems gemäß der KIVO und DIN 42001:2023 zu erklären.

Risikostufen im AI-Gesetz

Das AI-Gesetz klassifiziert KI-Systeme in mehrere Risikostufen, von denen jede spezifische Anforderungen zur Gewährleistung von Sicherheit, Transparenz und Compliance hat. Diese Kategorien sind

darauf ausgelegt, öffentliche Interessen wie Gesundheit, Sicherheit und grundlegende Rechte zu schützen und gleichzeitig Innovation zu fördern. Zur Einstufung finden sich mehr Infos unter X.

Verbotene KI-Praktiken

Die KIVO identifiziert bestimmte KI-Praktiken, die als untragbare Risiken angesehen werden und daher vollständig verboten sind. Diese Praktiken stehen in grundlegendem Widerspruch zu den in der EU-Gesetzgebung verankerten Werten und Rechten. Beispielsweise sind KI-Systeme, die darauf abzielen, menschliches Verhalten in einer Weise zu manipulieren, die Schaden verursacht, streng verboten. Dies könnte Nudging-Techniken umfassen, die Personen ohne deren bewusste Wahrnehmung beeinflussen sollen. Darüber hinaus sind Systeme, die eine soziale Bewertung durch Regierungen ermöglichen und zu ungerechter oder diskriminierender Behandlung basierend auf sozialem Verhalten oder Merkmalen führen, nicht erlaubt. Auch die Nutzung von KI-Systemen zur biometrischen Erkennung in öffentlichen Räumen durch Strafverfolgungsbehörden ist weitestgehend verboten.

Was ist zu tun? In Fällen, in denen KI-Systeme als verboten gelten, ist die sofortige Einstellung jeglicher laufender Nutzung erforderlich. Rechtliche Durchsetzungsmaßnahmen müssen ergriffen werden, um sicherzustellen, dass diese Systeme weder entwickelt noch innerhalb der Europäischen Union eingesetzt werden.

Hochrisiko-KI-Systeme

Hochrisiko-KI-Systeme sind solche, die erhebliche Risiken für Gesundheit, Sicherheit und grundlegende Rechte darstellen. Diese Systeme unterliegen strengen regulatorischen Anforderungen, um sicherzustellen, dass sie sicher und ethisch betrieben werden (wird auch im vorherigen Abschnitt erklärt). Beispiele für Hochrisiko-KI-Systeme sind solche, die in kritischen Infrastrukturen verwendet werden, wie etwa KI-Systeme zur Verwaltung von essenziellen Dienstleistungen wie Wasser, Energie und Transport. Im Bereich Bildung und berufliche Ausbildung bestimmen Hochrisiko-KI-Systeme den Zugang zu Bildungs- und Ausbildungsmöglichkeiten und beeinflussen damit wesentliche Lebensentscheidungen für Einzelpersonen. Ähnlich verhält es sich im Beschäftigungsbereich, wo KI-Systeme in Rekrutierungsprozessen, Leistungsbewertungen und Entscheidungsfindungen als Hochrisiko gelten, da sie tiefgreifende Auswirkungen auf die Lebensgrundlagen der Menschen haben.

Hochrisiko-KI-Systeme erstrecken sich auch auf essenzielle private und öffentliche Dienstleistungen, bei denen sie den Zugang zu Finanzdienstleistungen, Sozialleistungen und Versorgungsleistungen bestimmen. Im Kontext der Strafverfolgung und Migration spielen biometrische Identifikationssysteme und KI-Systeme zur Risikobewertung eine entscheidende Rolle und werden daher als Hochrisiko eingestuft.

Was ist zu tun? Die Anforderungen an Hochrisiko-KI-Systeme sind in den Artikeln 8 bis 15 der KIVO ausführlich dargelegt. Diese Artikel beschreiben die strengen Standards und Prozesse, die eingehalten werden müssen, um sicherzustellen, dass die Systeme verantwortungsbewusst entwickelt, eingesetzt und betrieben werden.

Artikel 8

Gemäß Artikel 8 müssen Hochrisiko-KI-Systeme strengen Standards entsprechen, die den beabsichtigten Zweck des Systems und den Stand der Technik berücksichtigen. Die Einhaltung umfasst umfassende Tests, Dokumentations- und Überwachungsprozesse, um sicherzustellen, dass das System sicher und effektiv ist.

Artikel 9

Artikel 9 schreibt vor, dass Anbieter ein umfassendes Risikomanagementsystem implementieren. Dieses System muss kontinuierlich und iterativ sein, wobei Risiken während des gesamten Lebenszyklus des Systems identifiziert, analysiert, bewertet und gemindert werden. Anbieter müssen gezielte Risikomanagementmaßnahmen ergreifen und umfassende Tests durchführen, um die fortlaufende Einhaltung und Leistung des Systems zu gewährleisten. Darauf wird am Ende dieses Abschnitts nochmal eingegangen.

Artikel 10

Artikel 10 betont die Bedeutung der Datenqualität für Hochrisiko-KI-Systeme. Anbieter müssen Datensätze verwenden, die relevant, repräsentativ, fehlerfrei und vollständig sind. Eine ordnungsgemäße Datenverwaltung ist entscheidend, um die Zuverlässigkeit und Fairness des KI-Systems zu gewährleisten.

Artikel 11

Umfassende technische Dokumentation, wie in Artikel 11 gefordert, ist unerlässlich, um die Einhaltung zu demonstrieren. Dies umfasst detaillierte Informationen zum Design des Systems, zu den Entwicklungsprozessen und zu den durchgeführten Tests, um Transparenz und Verantwortlichkeit zu gewährleisten.

Artikel 12

Artikel 12 beschreibt die Pflichten zur Aufbewahrung von Aufzeichnungen und verlangt, dass Betreiber detaillierte Aufzeichnungen über den Betrieb und die Leistung des Systems führen. Dies ist entscheidend für die Prüfungsfähigkeit und Verantwortlichkeit, um sicherzustellen, dass die Funktion des Systems überprüft werden kann und eventuelle Probleme umgehend behoben werden können.

Artikel 13

Die Transparenz und Bereitstellung von Informationen für Benutzer, wie in Artikel 13 beschrieben, erfordert, dass Benutzer klare und verständliche Informationen über das KI-System erhalten. Dies umfasst detaillierte Nutzungshinweise und Informationen zu den Fähigkeiten und Einschränkungen des Systems, sodass Benutzer fundierte Entscheidungen treffen können. Hier kann auch an der Vereinfachung von z.B. Interfaces gearbeitet werden, um Artikel 13 besser zu erfüllen.

Artikel 14

Artikel 14 unterstreicht die Bedeutung von menschlicher Aufsicht, auch human oversight im Englischen. Eine angemessene menschliche Überwachung stellt sicher, dass die Leistung des KI-Systems kontinuierlich gemonitored wird und dass menschliche Bediener eingreifen können, wenn dies erforderlich ist, um Schäden zu verhindern oder Risiken zu mindern. Diese müssen aber auch angemessen in die Lage versetzt werden, Systeme zu überwachen.

Artikel 15

Schließlich schreibt Artikel 15 vor, dass Hochrisiko-KI-Systeme hohe Genauigkeit, Robustheit und Cybersicherheit gewährleisten müssen. Dies umfasst die Gestaltung des Systems, um genau zu funktionieren, robust gegenüber Fehlern zu sein und gegen Cyber-Bedrohungen geschützt zu sein, um die Integrität und Sicherheit des Systems während seines gesamten Lebenszyklus zu gewährleisten.

Begrenzte Risikostufe

Begrenzte Risikostufe-KI-Systeme unterliegen spezifischen Transparenzverpflichtungen, insbesondere in Situationen, in denen es für Benutzer nicht offensichtlich ist, dass sie mit einem KI-System interagieren. Transparenz ist entscheidend, um sicherzustellen, dass Benutzer sich bewusst sind und fundierte Entscheidungen über ihre Interaktionen mit solchen Systemen treffen können.

Was ist zu tun? Im Kontext von begrenzten Risikostufe-KI-Systemen müssen Anbieter Benutzer klar darüber informieren, wenn sie mit einem KI-System interagieren, es sei denn, dies ist aus dem Kontext offensichtlich. Die Sicherstellung der Transparenz und Erklärbarkeit der Funktionsweise des Systems ist der Schlüssel zum Aufbau von Vertrauen und zur Ermöglichung der Benutzer, das Betrieb und die Entscheidungen des KI-Systems zu verstehen.

Minimale Risikostufe

Minimale Risikostufe-KI-Systeme umfassen Anwendungen wie Spamfilter oder KI-gestützte Videospiele, die nur geringe Risiken für Benutzer darstellen. Diese Systeme erfordern keine zusätzlichen spezifischen Verpflichtungen gemäß der KIVO, müssen jedoch die bestehenden Gesetze und Vorschriften einhalten.

Was ist zu tun? Für minimale Risikostufe-KI-Systeme ist es entscheidend, die allgemeinen gesetzlichen Anforderungen einzuhalten, um sicherzustellen, dass die Systeme im Rahmen der bestehenden Gesetze betrieben werden und keine unangemessenen Risiken für Benutzer darstellen.

3. KI-Managementsystem laut KIVO und DIN 42001:2023

Ein KI-Managementsystem, wie in der KIVO definiert, umfasst umfassende Prozesse und Verfahren, um sicherzustellen, dass KI-Systeme die regulatorischen Anforderungen erfüllen, Risiken managen und hohe Leistungsstandards aufrechterhalten. Dieses System sollte in die allgemeinen Managementprozesse der Organisation integriert werden.

Das KI-Managementsystem muss einen systematischen Ansatz zur Verwaltung aller KI-bezogenen Aktivitäten annehmen, um kontinuierliche Verbesserungen und Aktualisierungen der KI-Prozesse zu gewährleisten und sich an neue Entwicklungen anzupassen, um fortlaufende Compliance sicherzustellen. Es ist entscheidend, Rollen und Verantwortlichkeiten innerhalb der Organisation klar zu definieren, um eine Kultur der Verantwortung und Rechenschaftspflicht zu fördern.

Der DIN 42001:2023-Standard bietet Leitlinien zur Implementierung eines KI-Managementsystems und betont die Notwendigkeit eines systematischen und kontinuierlichen Ansatzes zur Verwaltung von KI-Technologien. Der Fokus liegt darauf, das KI-Managementsystem in die bestehenden Managementprozesse der Organisation zu integrieren, um sicherzustellen, dass KI-Systeme ethisch und verantwortungsvoll entwickelt, eingesetzt und genutzt werden.

Zusammenfassend lässt sich sagen, dass die KIVO einen umfassenden Rahmen für die Entwicklung, den Einsatz und die Nutzung von KI-Systemen schafft, diese in verschiedene Risikostufen einteilt und spezifische Anforderungen für jede Kategorie festlegt. Insbesondere Hochrisiko-KI-Systeme unterliegen strengen regulatorischen Anforderungen, um ihre Sicherheit und ethische Nutzung zu gewährleisten. Durch die Implementierung eines KI-Managementsystems gemäß der KIVO und DIN 42001:2023 können Organisationen effektiv KI-bezogene Aktivitäten verwalten, Compliance sicherstellen, verantwortungsvollen Einsatz fördern und Innovation in KI-Technologien vorantreiben.

High-Level Expert Group

Kursübersicht > [EU AI Act](#)

Es wird die HLEG betrachtet, was deren Ziele sind und welchen Einfluss sie haben.

1. Die Rolle der High-Level Expert Group für Trustworthy AI

Die High-Level Expert Group on Artificial Intelligence und ihre Ziele

Die High-Level Expert Group on Artificial Intelligence (HLEG) wurde im Juni 2018 von der Europäischen Kommission ins Leben gerufen. Diese unabhängige Expertengruppe besteht aus einer vielfältigen Auswahl von Fachleuten aus Wissenschaft, Industrie und Zivilgesellschaft. Ihr Hauptziel ist es, die Entwicklung und Implementierung von Künstlicher Intelligenz (KI) in Europa zu fördern und sicherzustellen, dass diese

Technologien im Einklang mit europäischen Werten und Grundrechten stehen (European Commission, 2020a).

Die HLEG verfolgt mehrere wesentliche Ziele. Erstens soll durch die Definition von Anforderungen für vertrauenswürdige KI das Vertrauen in diese Technologien gestärkt werden. Dies ist besonders wichtig, da die Akzeptanz von KI in der Gesellschaft davon abhängt, dass die Menschen den Systemen vertrauen und sich sicher fühlen. Zweitens sollen ethische Prinzipien und Grundrechte geschützt werden. Dies bedeutet, dass KI-Systeme nicht nur technisch einwandfrei sein müssen, sondern auch moralische und ethische Standards einhalten müssen. Drittens soll die technische Exzellenz gefördert werden. KI-Systeme müssen robust und sicher sein, um Risiken und potenzielle Schäden zu minimieren. Viertens soll die Wettbewerbsfähigkeit europäischer Unternehmen gestärkt werden. Durch die Entwicklung innovativer und vertrauenswürdiger KI-Lösungen können sich europäische Unternehmen im globalen Wettbewerb behaupten (European Commission, 2020b).

Ein zentrales Dokument der HLEG sind die „Ethics Guidelines for Trustworthy AI“, die im April 2019 veröffentlicht wurden. Diese Leitlinien definieren, was vertrauenswürdige KI ausmacht und welche Anforderungen an solche Systeme gestellt werden sollten. Vertrauenswürdige KI basiert auf drei Hauptkomponenten: Gesetzeskonformität, Ethik und Robustheit. Gesetzeskonformität bedeutet, dass KI-Systeme alle relevanten Gesetze und Vorschriften einhalten müssen. Ethik umfasst die Respektierung ethischer Prinzipien und Werte, während Robustheit sicherstellt, dass die Systeme technisch und sozial robust sind, um unabsichtliche Schäden zu vermeiden (High-Level Expert Group on Artificial Intelligence, 2019).

2. Die sieben Anforderungen der HLEG

Um diese Ziele zu erreichen, hat die HLEG sieben zentrale Anforderungen definiert:

1. Menschliche Autonomie und Aufsicht

KI-Systeme sollten die Entscheidungsfreiheit und -kompetenz der Menschen unterstützen und nicht untergraben. Dies bedeutet, dass KI als Unterstützung für menschliche Entscheidungen dienen sollte und Mechanismen zur menschlichen Aufsicht und Kontrolle der Systeme implementiert werden müssen. Beispielsweise könnten Systeme entwickelt werden, die klare Hinweise geben, wann eine menschliche Überprüfung erforderlich ist, oder die es den Nutzern ermöglichen, Entscheidungen der KI-Systeme zu hinterfragen und zu überstimmen.

2. Technische Robustheit und Sicherheit

Zuverlässigkeit, Sicherheit und Robustheit der KI-Systeme sind entscheidend, um die Integrität und Vertrauenswürdigkeit der Technologien zu gewährleisten. Technische Robustheit erfordert die Fähigkeit der Systeme, in verschiedenen Situationen und unter unterschiedlichen Bedingungen zuverlässig zu funktionieren. Sicherheit bedeutet, dass Systeme vor Angriffen geschützt und Ausfälle minimiert werden müssen. Dies schließt auch die Notwendigkeit ein, dass KI-Systeme regelmäßig getestet und aktualisiert werden, um sicherzustellen, dass sie sicher und effektiv bleiben (European

Commission, 2020a; High-Level Expert Group on Artificial Intelligence, 2019).

3. Privatsphäre und Datenmanagement

Der Schutz der Privatsphäre und eine verantwortungsbewusste Datenverwaltung sind essenziell. KI-Systeme müssen so gestaltet sein, dass sie die Privatsphäre der Nutzer respektieren und schützen. Dies umfasst Maßnahmen zum Datenschutz, zur Qualität und Integrität der Daten sowie zu transparenten Zugriffs- und Verarbeitungsprotokollen. Beispielsweise sollten Daten, die von KI-Systemen gesammelt werden, anonymisiert oder pseudonymisiert werden, um die Privatsphäre der Nutzer zu schützen (European Commission, 2020a).

4. Transparenz

Nachvollziehbarkeit, Erklärbarkeit und klare Kommunikation der KI-Systeme sind entscheidend, um Vertrauen in diese Technologien zu schaffen. Transparenz bedeutet, dass die Entscheidungen und Funktionsweisen der KI-Systeme verständlich und zugänglich erklärt werden müssen. Dies ermöglicht es den Nutzern, die Funktionsweise der Systeme zu verstehen und ihre Entscheidungen zu hinterfragen. Erklärbarkeit bezieht sich darauf, dass die Prozesse, die zu einer bestimmten Entscheidung führen, klar und nachvollziehbar sind (High-Level Expert Group on Artificial Intelligence, 2019).

5. Vielfalt, Nichtdiskriminierung und Fairness

KI-Systeme sollen inklusiv gestaltet sein und dürfen keine diskriminierenden Auswirkungen haben. Dies erfordert die Beseitigung von Verzerrungen in den Daten und Modellen sowie die Berücksichtigung

aller Nutzergruppen. Beispielsweise sollten KI-Systeme so entwickelt werden, dass sie alle Nutzer unabhängig von deren ethnischer Herkunft, Geschlecht, Alter oder anderen persönlichen Merkmalen gleich behandeln. Dies könnte durch die Implementierung von Mechanismen zur Überprüfung und Korrektur von Verzerrungen in den Daten und Algorithmen erreicht werden (European Commission, 2020a).

6. Gesellschaftliches und ökologisches Wohlergehen

Die Auswirkungen von KI auf Gesellschaft und Umwelt müssen berücksichtigt werden. Dazu gehört die Förderung des gesellschaftlichen Wohlergehens und die Minimierung negativer Umweltauswirkungen. KI-Systeme sollten so gestaltet sein, dass sie positive soziale und ökologische Auswirkungen haben. Beispielsweise könnten KI-Systeme entwickelt werden, die Energieeffizienz verbessern oder zur Lösung sozialer Probleme beitragen (High-Level Expert Group on Artificial Intelligence, 2019).

7. Rechenschaftspflicht

Es muss klare Verantwortlichkeiten und Mechanismen zur Überprüfung und Rechenschaftspflicht geben. Dies umfasst die Möglichkeit zur Überprüfung der Systeme durch unabhängige Dritte sowie die Implementierung von Mechanismen zur Korrektur von Fehlern und zur Rechenschaftspflicht derjenigen, die die Systeme entwickeln und einsetzen. Rechenschaftspflicht bedeutet auch, dass klare Verfahren und Standards zur Überprüfung und Bewertung der KI-Systeme etabliert werden müssen (European Commission, 2020a; High-Level Expert Group on Artificial Intelligence, 2019).

3. Unterschiede zwischen dem AI Act und den Anforderungen der HLEG

Die Arbeit der High-Level Expert Group on AI und der EU AI Act stellen wesentliche Schritte dar, um sicherzustellen, dass KI-Systeme im Einklang mit europäischen Werten und Grundrechten entwickelt und eingesetzt werden. Während die HLEG einen breiteren, ethisch orientierten Ansatz verfolgt, legt der EU AI Act einen stärkeren Fokus auf die Regulierung basierend auf dem Risiko, das KI-Systeme darstellen.

Ein wesentlicher Unterschied liegt in der Herangehensweise: Der EU AI Act kategorisiert KI-Systeme basierend auf dem Risiko, das sie darstellen. Es gibt verbotene Praktiken, Hochrisiko-Systeme, Systeme mit eingeschränktem Risiko und Systeme mit minimalem Risiko. Diese risikobasierte Herangehensweise bestimmt, welche Anforderungen an die jeweiligen KI-Systeme gestellt werden. Im Gegensatz dazu definiert die HLEG allgemeine Anforderungen, die für alle KI-Systeme gelten sollten, unabhängig von ihrem spezifischen Risikopotenzial (European Commission, 2020b).

Besonders hervorzuheben sind die Themen Vielfalt und ökologische Wirkung, die von der HLEG stärker betont werden als vom EU AI Act. Die HLEG legt großen Wert auf die Inklusion aller Nutzergruppen und die Vermeidung von Diskriminierung. Dies erfordert Maßnahmen zur Beseitigung von Verzerrungen in den Daten und Modellen sowie die Berücksichtigung aller demografischen Gruppen. Der EU AI Act hingegen fokussiert sich stärker auf rechtliche und technische

Anforderungen und behandelt die Vielfalt weniger ausführlich (European Commission, 2020a).

Ebenso legt die HLEG besonderen Wert auf das gesellschaftliche und ökologische Wohlergehen. KI-Systeme sollten nicht nur technisch robust sein, sondern auch positive soziale und ökologische Auswirkungen haben. Dies schließt die Förderung von Energieeffizienz und die Minimierung negativer Umweltauswirkungen ein. Der EU AI Act hingegen betont stärker die Einhaltung rechtlicher Standards und die technische Sicherheit, während die ökologischen Aspekte weniger prominent behandelt werden (High-Level Expert Group on Artificial Intelligence, 2019).

4. Zusammenfassung der Kriterien und Ziele der HLEG

- Menschliche Autonomie und Aufsicht
- Technische Robustheit und Sicherheit
- Privatsphäre und Datenmanagement
- Transparenz
- Vielfalt, Nichtdiskriminierung und Fairness
- Gesellschaftliches und ökologisches Wohlergehen
- Rechenschaftspflicht

Die Arbeit der High-Level Expert Group on AI stellt einen wichtigen Schritt dar, um sicherzustellen, dass KI-Systeme im Einklang mit europäischen Werten und Grundrechten entwickelt und eingesetzt werden. Definierte Kriterien und Leitlinien dienen dabei als Grundlage

für die Entwicklung und Implementierung vertrauenswürdiger und ethisch verantwortungsvoller KI-Systeme. Die umfassende Betrachtung technischer, ethischer und sozialer Aspekte trägt dazu bei, dass KI nicht nur technologisch fortschrittlich, sondern auch sozial und ethisch verantwortungsvoll ist.

06

Quellen

Kursübersicht > [EU AI Act](#)

Literaturverzeichnis

- European Commission. (2020a). *The Ethics Guidelines for Trustworthy AI*. Abgerufen von <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- European Commission. (2020b). *The EU AI Act: Regulatory framework proposal for Artificial Intelligence*. Abgerufen von <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206>