

# 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?

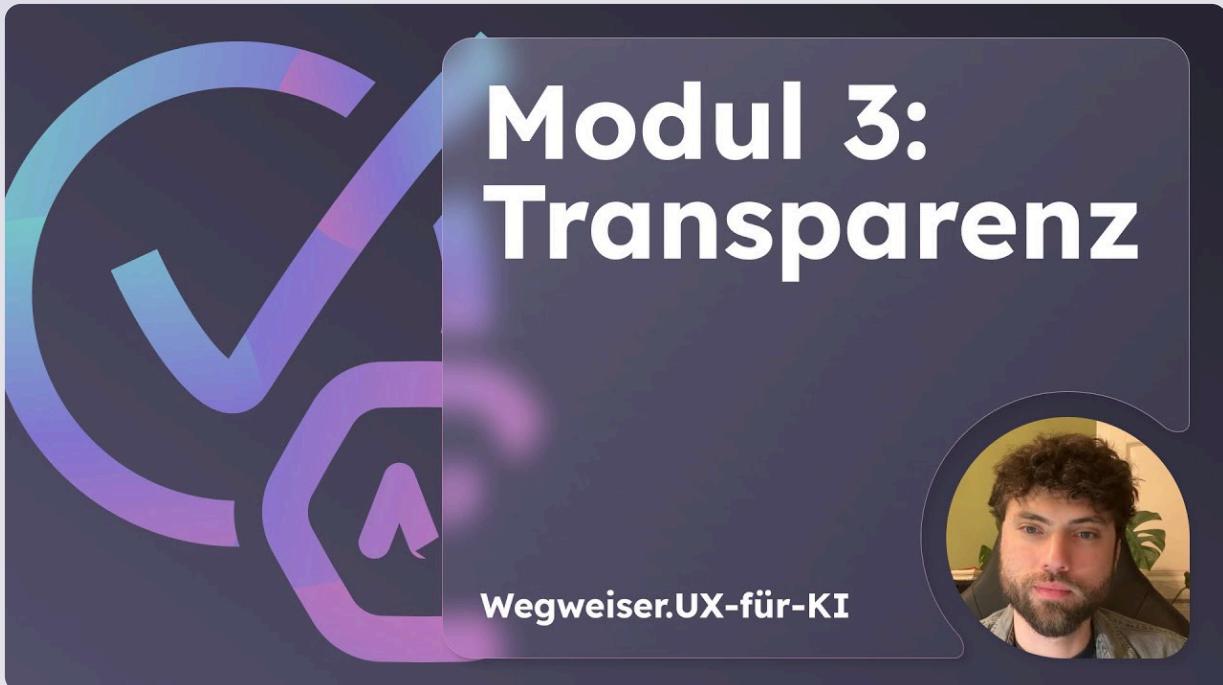


# Transparenz in KI-Systemen

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.

- **Beispiele**: 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 plausible 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](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.