Weisheitskompass und Hararis Warnung

Orientierungs- und Handlungskompetenz

unter Bedingungen algorithmischer Informationsmacht

Ein wissenschaftlicher Essay mit mathematischem Erklärmodell

© Norbert Rieser

Stand: 21. August 2026

Wissenschaftlicher Status:
erkenntnistheoretisch-handlungstheoretische Analyse;
das mathematische Modell dient der formalen Klärung
und erhebt keinen Anspruch auf empirische Validierung als Prognoseinstrument.

Inhaltsverzeichnis

Abstract

Mit der raschen Verbreitung generativer künstlicher Intelligenz rückt eine klassische erkenntnistheoretische Frage in den Alltag: Wie lässt sich verlässlich urteilen, wenn technische Systeme Informationen auswählen, verdichten, personalisieren und mit erheblicher sprachlicher Überzeugungskraft darbieten? Yuval Noah Harari hat diese Entwicklung in mehreren Werken zugespitzt. Homo Deus entwirft die Möglichkeit einer schrittweisen Verlagerung von Entscheidungsautorität auf datengetriebene Systeme; 21 Lessons for the 21st Century richtet den Blick auf die Überforderung durch Informationsfülle; Nexus untersucht das Verhältnis von Informationsnetzwerken, Macht und Wahrheit unter den Bedingungen nichtmenschlicher Intelligenz. Zwei öffentlich zugängliche Gespräche – die SRF-Sternstunde Philosophie im Umfeld von Homo Deus und ein späteres Langinterview bei The Diary of a CEO – verdeutlichen die zeitliche Entwicklung dieser Warnung.1

Hararis Aussagen werden im Folgenden als historisch informierte und gesellschaftstheoretisch weitreichende Problemhypothesen behandelt. Für empirische Teilfragen zieht die Analyse unabhängige Fachliteratur heran. Neuere Experimente weisen nach, dass große Sprachmodelle politische und gesellschaftliche Einstellungen messbar beeinflussen können; zugleich variieren die Wirkungen algorithmischer Feeds erheblich nach Plattform, Systemarchitektur und Untersuchungskontext. Die derzeitige Befundlage trägt daher weder ein Szenario allmächtiger algorithmischer Manipulation noch eine beruhigende Annahme grundsätzlicher Wirkungslosigkeit.2

Den Ordnungsrahmen liefert der Weisheitskompass. Seine veröffentlichte Fassung verbindet Erkennen und Deuten sowie Fühlen und Handeln mit Integration und einer Reifedimension. Die spätere mathematische Weiterentwicklung ergänzt getrennte Geltungsarten, partielle Beobachtbarkeit, probabilistische Wissensaktualisierung, Kausalitätsprüfung, normative Grenzen, Mehrkriterienentscheidungen, Robustheit, Informationswert, Reversibilität und systemische Rückkopplung. Aus dieser Verbindung entsteht ein Verständnis von Orientierungs- und Handlungskompetenz, das technische Assistenz zulässt und zugleich die Prüfung von Gründen, Werten, Folgen und Verantwortlichkeiten beim Menschen verankert. Der mathematische Anhang fasst diese Beziehung als gekoppelten Mensch-Algorithmus-Regelkreis und kennzeichnet ausdrücklich die Grenzen seiner Messbarkeit.3

1. Ausgangslage: Informationsmacht als Orientierungsfrage

Die öffentliche Auseinandersetzung mit künstlicher Intelligenz bewegt sich zwischen Faszination und Alarm. Auf der einen Seite stehen Systeme, die Texte zusammenfassen, Übersetzungen erstellen, Programmcode erzeugen, medizinische oder technische Informationen strukturieren und große Datenmengen in kurzer Zeit erschließen. Auf der anderen Seite wächst die Sorge vor massenhaft erzeugten Falschinformationen, personalisierter Überredung und einer kaum bemerkten Vorstrukturierung von Entscheidungen. Damit verschiebt sich der Schwerpunkt der Debatte: Im Zentrum steht zunehmend die

1 Zu Hararis Werkentwicklung: SRF, Sternstunde Philosophie, „Yuval Harari: Ein Historiker erzählt die Geschichte von morgen“: https://www.srf.ch/play/tv/sternstunde-philosophie/video/yuval-harari-ein-historiker-erzaehlt-die-geschichte-von-morgen?urn=urn%3Asrf%3Avideo%3A8b1ea2c8-0be4-44cb-82bc-6c3d7631ceba; Harari, Nexus, offizielle Werkseite: https://www.ynharari.com/book/nexus/

2 Bai et al. (2025), Nature Communications, LLM-generierte Persuasion: https://doi.org/10.1038/s41467-025-61345-5; Nyhan et al. (2023), Nature, Meta-Feed-Experiment: https://doi.org/10.1038/s41586-023-06297-w; Gauthier et al. (2026), Nature, X-Feed-Experiment: https://doi.org/10.1038/s41586-026-10098-2

3 Rieser, Der Weisheitskompass 2.0, öffentliche Fassung: https://assets.sta.io/site_media/u/files/2026/01/31/Weisheitskompass_2.0_Rev_ygOMHc9.pdf; ergänzend: Rieser (2026), Verantwortliches Handeln unter Unsicherheit, Manuskript, und Praxisanhang, auf Homepage veröffentlichtes Manuskript.

Frage, wie Menschen ihre Urteilskraft unter Bedingungen technisch vermittelter Informationsmacht bewahren und weiterentwickeln können.

Die beiden im Ausgangsmaterial sichtbaren Videoaufnahmen veranschaulichen zwei Etappen dieser Entwicklung. In der SRF-Sternstunde Philosophie aus dem Umfeld von Homo Deus erörtert Harari mit Barbara Bleisch eine Zukunft, in der Daten und Algorithmen menschliche Entscheidungen tiefgreifend verändern könnten. Das jüngere Gespräch bei The Diary of a CEO führt denselben Gedanken in eine Gegenwart, in der generative KI bereits als Gesprächspartner, Empfehlungssystem und Produzent politisch relevanter Inhalte auftritt.4

An dieser Stelle setzt der Weisheitskompass an. Information erweitert den Handlungsspielraum erst dann, wenn Menschen Relevanz, Geltungsanspruch und Tragweite einordnen können. Eine ungeordnete Informationsfülle kann Verwirrung vermehren; präzise Prognosen können Entscheidungen effizienter machen und zugleich den Raum eigenständiger Abwägung verengen, sofern normative Maßstäbe ungeklärt bleiben. Daraus ergibt sich die leitende Forschungsfrage: Welche Orientierungs- und Handlungskompetenz benötigen Menschen, wenn algorithmische Systeme zu aktiven Mitspielern ihrer Informations- und Entscheidungsprozesse werden?

2. Quellenbasis, Methode und urheberrechtliche Selbstbegrenzung

Die Untersuchung verbindet vier Quellengruppen. Hararis veröffentlichte Werke und offizielle Werkbeschreibungen – insbesondere Homo Deus, 21 Lessons for the 21st Century und Nexus – bilden die erste Gruppe; ausgewählte Medienauftritte dienen der zeitlichen und argumentativen Einordnung seiner Thesen. Die zweite Gruppe umfasst aktuelle empirische Forschung zu Persuasion durch große Sprachmodelle, Empfehlungsalgorithmen, politischer Meinungsbildung und menschlicher Autonomie. Internationale und europäische Governance-Dokumente bilden die dritte Ebene, weil Transparenz, menschliche Aufsicht und die Übersteuerbarkeit automatisierter Ausgaben inzwischen rechtlich und institutionell konkretisiert werden. Den vierten Bezugspunkt bilden die veröffentlichte Fassung des Weisheitskompasses 2.0 sowie zwei spätere Manuskripte zur mathematischen und praktischen Weiterentwicklung.56

Methodisch verlangt die Untersuchung eine konsequente Trennung der Geltungsansprüche. Hararis Zukunftsaussagen werden als historisch informierte Risikoszenarien und gesellschaftstheoretische Deutungen gelesen. Experimentelle Studien erlauben enger umrissene Aussagen über beobachtete Wirkungen unter bestimmten Bedingungen. Normative Folgerungen, etwa zur Bedeutung menschlicher Autonomie, benötigen darüber hinaus eigenständige Begründungen. Diese Staffelung verhindert, dass ein eindrucksvolles Szenario den Rang einer bereits bewiesenen Zukunft erhält oder ein einzelnes Experiment die Last einer umfassenden Kulturdiagnose tragen muss.

---

4 SRF-Video: https://www.srf.ch/play/tv/sternstunde-philosophie/video/yuval-harari-ein-historiker-erzaehlt-die-geschichte-von-morgen?urn=urn%3Asrf%3Avideo%3A8b1ea2c8-0be4-44cb-82bc-6c3d7631ceba; The Diary of a CEO, Gespräch mit Yuval Noah Harari vom 5. September 2024: https://www.youtube.com/watch?v=78YN1e8UXdM

5 Offizielle Werkseiten: Homo Deus: https://www.ynharari.com/book/homo-deus/; 21 Lessons for the 21st Century: https://www.ynharari.com/book/21-lessons-book/; Nexus: https://www.ynharari.com/book/nexus/

6 Rieser, Der Weisheitskompass 2.0: https://assets.sta.io/site_media/u/files/2026/01/31/Weisheitskompass_2.0_Rev_ygOMHc9.pdf; Rieser (2026), Verantwortliches Handeln unter Unsicherheit und Praxisanhang, auf Homepage veröffentlichte Manuskripte.

Die urheberrechtliche Selbstbegrenzung folgt demselben Prinzip wissenschaftlicher Redlichkeit. Weder die bereitgestellten Bildschirmfotos noch längere Passagen aus Videos, Büchern oder Transkripten werden reproduziert. Die Darstellung arbeitet mit eigenständigen Paraphrasen; kurze Begriffe erscheinen nur dort, wo ihre analytische Funktion selbst zur Diskussion steht. Sämtliche Internetadressen dienen dem Quellen- und Fundstellennachweis. Aus ihrer Verlinkung werden keinerlei weitergehende Nutzungs- oder Verwertungsrechte abgeleitet.

3. Hararis Warnung in ihrer Entwicklung

3.1 Homo Deus: Die mögliche Verlagerung von Entscheidungsautorität

In Homo Deus verdichtet Harari die Digitalisierung zu einer Machtfrage. Die offizielle Buchdarstellung entwirft provokativ die Möglichkeit, dass demokratische und marktwirtschaftliche Entscheidungsordnungen unter Druck geraten könnten, sobald große Plattformen menschliches Verhalten verlässlicher prognostizieren als die Betroffenen selbst und sich Entscheidungsautorität schrittweise zu vernetzten Algorithmen verlagert. Die Zuspitzung erfüllt erkennbar die Funktion einer Warnung: Sie skizziert einen möglichen Entwicklungspfad aus Datenkonzentration, Verhaltensprognose und wachsender Delegation, ohne einen empirisch gesicherten Endzustand zu behaupten.7

Aufmerksamkeit verdient der erkenntnistheoretische Kern dieser Warnung. Für eine erfolgreiche Verhaltensprognose muss ein System das innere Erleben eines Menschen keineswegs vollständig verstehen; in vielen praktischen Situationen genügt eine statistisch hinreichend verlässliche Vorhersage. Wer regelmäßig die passende Route, einen interessierenden Film oder eine vorteilhafte Kaufoption vorgeschlagen bekommt, entwickelt nachvollziehbare Gründe, algorithmischen Empfehlungen zu folgen. Aus vielen kleinen Delegationen kann sich jedoch ein Gewöhnungseffekt bilden: Die eigene Urteilsarbeit verliert Übung, während die technische Empfehlung zunehmend den Charakter des Selbstverständlichen annimmt.

3.2 21 Lessons: Klarheit in der Informationsüberflutung

21 Lessons for the 21st Century verlagert den Akzent von der Datensammlung auf die Struktur der Informationsumgebung. Hararis offizielle Beschreibung eröffnet mit der These, dass Klarheit in einer von irrelevanten Informationen überfluteten Welt zu einer Machtressource werde; moderne Zensur könne ebenso durch Desinformation und Ablenkung wirken. Für den Weisheitskompass berührt dies einen zentralen Gedanken: Verfügbare Information gewinnt erst durch die Fähigkeit zur Unterscheidung von Relevanz, Geltungsart, Zuverlässigkeit und praktischer Bedeutung orientierende Kraft.8

Für Bildung hat diese Verschiebung erhebliche Folgen. Klassische Informationskompetenz konzentrierte sich lange auf Zugang und Recherche. In einer Umgebung nahezu unbegrenzter Verfügbarkeit gewinnt die Auswahl- und Prüfungskompetenz an Gewicht. Wer tausend Antworten erhält, braucht Kriterien dafür, welche davon eine nähere Prüfung verdienen, welche verworfen werden können und welche in eine Entscheidung eingehen sollen. Generative KI verschärft diese Aufgabe, weil sprachliche Qualität und sachliche Zuverlässigkeit auseinanderfallen können: Eine flüssige und plausible Antwort mag korrekt, unvollständig oder falsch ausfallen; ihr Stil liefert dafür keinen verlässlichen Indikator.

7 Harari, Homo Deus, offizielle Werkseite mit den zugespitzten Zukunftsthesen: https://www.ynharari.com/book/homo-deus/

8 Harari, 21 Lessons for the 21st Century, offizielle Werkseite: https://www.ynharari.com/book/21-lessons-book/

3.3 Nexus: Informationsnetzwerke, Wahrheit und nichtmenschliche Akteure

Mit Nexus weitet Harari seine Perspektive zu einer Geschichte der Informationsnetzwerke. Die offizielle Werkbeschreibung rückt Spannungen zwischen Information und Wahrheit, Bürokratie und Mythologie sowie Weisheit und Macht in den Vordergrund. Zugleich warnt sie vor der Möglichkeit, dass KI eigenständig Inhalte erzeugt und dadurch neue Netze der Täuschung stabilisieren könnte. Für die wissenschaftliche Einordnung besonders wichtig bleibt der ausdrücklich offene Ausgang: Gesellschaftliche Entscheidungen können Entwicklungspfade verändern.9

Diese Offenheit bewahrt die Argumentation vor technologischem Fatalismus. Wer technische Entwicklung als unabwendbares Schicksal auffasst, verkleinert den Raum politischer und persönlicher Verantwortung. Hararis spätere Position eröffnet dagegen eine Gestaltungsfrage: Welche institutionellen Regeln, Bildungsformen und individuellen Kompetenzen ermöglichen eine produktive Nutzung neuer Systeme, ohne vermeidbare Abhängigkeiten und Machtasymmetrien zu verstärken?

4. Information, Wahrheit und Wissen: begriffliche Trennlinien

Digitale Systeme operieren mit Information. Wahrheit betrifft die Beziehung einer Aussage zu ihrem Gegenstand; Wissen verlangt darüber hinaus eine hinreichende Begründung. Diese Staffelung schützt vor einer folgenreichen Verwechslung. Auch eine sehr große Datenmenge kann ein verzerrtes Gesamtbild hervorbringen, wenn Auswahl, Gewichtung oder Kontext fehlen. Umgekehrt kann eine knappe Information außerordentlich orientierungswirksam werden, sofern sie zuverlässig, relevant und sachgerecht eingeordnet wurde.

Bei generativer KI tritt eine weitere Ebene hinzu: sprachliche Kohärenz. Große Sprachmodelle erzeugen Texte, die strukturiert, sicher und argumentativ geschlossen wirken können. Die sprachliche Oberfläche vermittelt damit leicht mehr Gewissheit, als der epistemische Unterbau rechtfertigt. Orientierungskompetenz verlangt folglich, rhetorische Qualität und Evidenzqualität getrennt zu beurteilen. Die Frage nach der Plausibilität einer Formulierung gewinnt erst im Zusammenspiel mit der Prüfung ihrer Quellen, Voraussetzungen und möglichen Gegenbefunde wissenschaftlichen Wert.

Für diese Differenzierung stellt der Weisheitskompass ein präzises Vokabular bereit. Seine mathematische Weiterentwicklung unterscheidet deduktiv-logische, empirisch-probabilistische, interpretative, normative sowie Glaubens- und Transzendenzaussagen. Jede dieser Geltungsarten verlangt eigene Prüfregeln. Ein Wahrscheinlichkeitsurteil kann daher keine moralische Pflicht erzeugen; eine Wertentscheidung darf sich auf Tatsachen stützen, erhält ihre normative Verbindlichkeit jedoch aus zusätzlichen Gründen.10

5. Empfehlungsalgorithmen und die Rückwirkung der Auswahl

Empfehlungssysteme helfen, ein kaum überschaubares Angebot zu ordnen. Sie gewichten Nachrichten, Videos, Musik, Produkte oder Kontakte anhand von Nutzerdaten und Ähnlichkeitsmustern zwischen Personen und Inhalten. Diese Funktion kann Autonomie unterstützen, indem sie Entscheidungsaufwand verringert und relevante Optionen sichtbar macht. Zugleich verweist die ethische Literatur auf Risiken, die

9 Harari, Nexus, offizielle Werkseite: https://www.ynharari.com/book/nexus/; Quellenapparat zu Nexus: https://www.ynharari.com/wp-content/uploads/2024/11/Notes-NEXUS-Yuval-Noah-Harari-September-2024.pdf

10 Rieser (2026), Verantwortliches Handeln unter Unsicherheit, Kap. 4–11; wissenschaftlicher Status des Modells dort ausdrücklich als formale Meta- und Entscheidungsarchitektur begrenzt.

von Manipulation und Identitätsprägung über eingeschränkte Vielfalt bis zu mangelnder Transparenz und Beeinträchtigungen kritischen Denkens reichen.11

Die systemisch folgenreichste Eigenschaft solcher Systeme zeigt sich in der Rückkopplung. Eine Plattform beobachtet Verhalten, leitet daraus Präferenzen ab, verändert anschließend die Auswahl und gewinnt neue Verhaltensdaten aus der bereits veränderten Umgebung. Prognose und Intervention greifen dadurch ineinander. Die Plattform erfasst vorhandene Präferenzen und wirkt zugleich an jenen Bedingungen mit, unter denen künftige Präferenzen entstehen können.

Aus dieser Rückkopplung folgt noch keine pauschale Manipulationsdiagnose. Menschen verfügen über Gegenquellen, soziale Beziehungen, Gewohnheiten und eigene Überzeugungen; Plattformen unterscheiden sich zudem deutlich in Zielsetzung und technischer Architektur. Wissenschaftlich tragfähig bleibt deshalb eine konditionale Aussage: Personalisierte Auswahl kann Wahrnehmung und Verhalten beeinflussen, wobei Stärke, Richtung und Dauer der Wirkung vom jeweiligen System, vom Inhalt, von der Nutzergruppe und vom sozialen Kontext abhängen.

6. Empirische Befunde: Reichweite und Grenzen

6.1 LLM-generierte Persuasion

Neuere Experimente belegen, dass große Sprachmodelle persuasive Kommunikation erzeugen können. Bai und Kolleginnen beziehungsweise Kollegen untersuchten in drei vorregistrierten Studien mit insgesamt 4.829 Teilnehmenden politische und gesellschaftliche Themen. LLM-generierte Botschaften bewirkten gegenüber neutralen Kontrolltexten messbare Einstellungsverschiebungen; ihre Wirkung bewegte sich insgesamt in einer ähnlichen Größenordnung wie bei Texten von Laien. Damit lässt sich eine reale Fähigkeit zur Überredung empirisch stützen. Aussagen über allgegenwärtige oder gar unwiderstehliche Manipulation reichen über diesen Befund weit hinaus.12

Salvi und Kollegen erweiterten diese Fragestellung durch kurze Debatten mit GPT-4. Eine Personalisierung anhand soziodemografischer Angaben erhöhte die Überzeugungswirkung in bestimmten Konstellationen. Hararis Sorge vor individualisierter Beeinflussung erhält dadurch eine empirische Teilstütze. Für Aussagen über reale Wahlentscheidungen, langfristige Identitätsbildung oder komplexe gesellschaftliche Konflikte bleibt Zurückhaltung geboten, weil kontrollierte Experimente stets nur Ausschnitte dieser Wirklichkeit abbilden.13

6.2 Algorithmische Feeds und politische Einstellungen

Bei sozialen Medien fällt die Befundlage weniger einheitlich aus. Die groß angelegte Facebook- und Instagram-Wahlstudie zur US-Wahl 2020 zeigte eine hohe Präsenz gleichgesinnter Quellen im Feed. Die experimentelle Verringerung solcher Inhalte führte im untersuchten Zeitraum jedoch zu keiner entsprechenden Abnahme politischer Polarisierung. Damit verliert die verbreitete Erwartung, ein geringerer

11 Milano, Taddeo & Floridi (2020), „Recommender systems and their ethical challenges“, AI & Society 35, 957–967: https://doi.org/10.1007/s00146-020-00950-y; Bonicalzi, De Caro & Giovanola (2023): https://doi.org/10.1007/s11245-023-09922-5

12 Bai, H. et al. (2025): „LLM-generated messages can persuade humans on policy issues“, Nature Communications 16, 6037: https://doi.org/10.1038/s41467-025-61345-5

13 Salvi, F. et al. (2025): „On the conversational persuasiveness of GPT-4“, Nature Human Behaviour 9, 1645–1653: https://doi.org/10.1038/s41562-025-02194-6

Anteil gleichgesinnter Inhalte führe automatisch zu gemäßigteren Einstellungen, ihre Allgemeingültigkeit.14

Ein 2026 veröffentlichtes Feldexperiment auf X gelangte zu einem anderen Ergebnis. Der Wechsel von einem chronologischen zu einem algorithmischen Feed steigerte die Nutzung und verschob mehrere politische Einstellungen der untersuchten US-Nutzer in konservativere Richtung; bei affektiver Polarisierung und selbstberichteter Parteibindung fanden die Forschenden zugleich keine signifikanten Veränderungen. Der Vergleich beider Studien zeigt, wie stark algorithmische Effekte vom konkreten Plattformdesign und von der jeweils betrachteten politischen Dimension abhängen.15

6.3 Zwischen Alarmismus und vorschneller Entwarnung

Die empirische Befundlage entzieht sich einfachen Erzählungen. Algorithmen entfalten weder eine einheitliche politische Wirkung noch bleiben sie grundsätzlich folgenlos. Große Sprachmodelle können überzeugen, Personalisierung kann diese Wirkung in bestimmten Konstellationen verstärken, und einzelne Feed-Algorithmen verändern nachweisbar Exposition und Einstellungen. Andere Experimente markieren zugleich deutliche Grenzen. Wissenschaftlich angemessen erscheint daher eine Sprache abgestufter Wahrscheinlichkeiten, die den Geltungsbereich jeder Studie offenlegt.

Hier gewinnt der Weisheitskompass analytische Schärfe. Revisionsfähigkeit schützt vor zwei spiegelbildlichen Verkürzungen: technologischem Fatalismus und technologischem Naivismus. Beide verwandeln Unsicherheit vorschnell in Gewissheit. Verantwortliche Orientierung hält mehrere plausible Modelle offen und bevorzugt Handlungen, die unter unterschiedlichen Annahmen tragfähig bleiben.

7. Autonomie unter algorithmisch vermittelten Bedingungen

Autonomie umfasst mehr als die formale Wahl zwischen verfügbaren Optionen. Philosophische Analysen von Empfehlungssystemen beziehen auch die Fähigkeit ein, eigene Gründe zu prüfen, Entscheidungen reflektiert zu bejahen und auf relevante Gegenargumente reagieren zu können. Bonicalzi, De Caro und Giovanola benennen drei besonders sensible Bereiche: Manipulation und Täuschung, die Mitprägung persönlicher Identität sowie Folgen für Wissen und kritisches Denken. Zugleich würdigen sie die entlastende Funktion guter Empfehlungen in komplexen Informationsumgebungen.16

Eine 2026 veröffentlichte Literaturübersicht zur Beziehung zwischen KI und menschlicher Autonomie bestätigt die Breite dieser Fragestellung. KI kann menschliche Tätigkeit vermitteln, indem sie Entscheidungen und Verhalten beeinflusst; sie kann Tätigkeiten auch automatisieren und dadurch die Bedingungen autonomen Handelns verändern. Die angemessene Bewertung hängt vom konkreten System, vom Anwendungskontext und vom zugrunde gelegten Autonomieverständnis ab.17

---

14 Nyhan, B. et al. (2023): „Like-minded sources on Facebook are prevalent but not polarizing“, Nature 620, 137–144: https://doi.org/10.1038/s41586-023-06297-w

15 Gauthier, G.; Hodler, R.; Widmer, P.; Zhuravskaya, E. (2026): „The political effects of X’s feed algorithm“, Nature 652, 416–423: https://doi.org/10.1038/s41586-026-10098-2

16 Bonicalzi, S.; De Caro, M.; Giovanola, B. (2023): „Artificial Intelligence and Autonomy: On the Ethical Dimension of Recommender Systems“, Topoi 42, 819–832: https://doi.org/10.1007/s11245-023-09922-5

17 Catena, E. (2026): „AI and human autonomy: a literature review“, AI and Ethics 6, Article 126: https://doi.org/10.1007/s43681-025-00958-4

Für den Weisheitskompass folgt eine präzise Konsequenz. Autonomie setzt keine unrealistische Abwesenheit von Einfluss voraus; ausschlaggebend wird vielmehr die Qualität der Einflussbeziehung. Erforderlich sind hinreichende Transparenz, Zugang zu Gegenperspektiven, die reale Möglichkeit zur Ablehnung einer Empfehlung und die Fähigkeit, Gründe eigenständig zu gewichten. Unter solchen Bedingungen kann technische Assistenz den Handlungsspielraum erweitern. Wo undurchsichtige Steuerung Kritik erschwert oder Alternativen systematisch verengt, nimmt dagegen die Gefahr der Fremdorientierung zu.

8. Der Weisheitskompass als Architektur verantwortlicher Orientierung

8.1 Grundmodell: Erkennen – Deuten, Fühlen – Handeln

Die veröffentlichte Fassung des Weisheitskompasses 2.0 ordnet Orientierung entlang zweier Achsen: Erkennen und Deuten sowie Fühlen und Handeln. Integration bildet das Zentrum; eine Reifedimension beschreibt den Weg von reaktiver über reflexive zu integrierter Orientierung. Damit trägt das Modell der Erfahrung Rechnung, dass menschliche Entscheidungen aus einem Zusammenspiel von Wahrnehmung, Bedeutung, Gefühl, Wertung und praktischer Konsequenz hervorgehen.18

Auf künstliche Intelligenz übertragen erhält jede Dimension eine eigene Prüfaufgabe. Erkennen richtet den Blick auf Daten, Quellen und empirische Tragfähigkeit. Deuten erschließt die Bedeutung einer Empfehlung im konkreten Lebens- und Gesellschaftskontext. Fühlen nimmt wahr, welche Ängste, Hoffnungen, Kränkungen oder Zugehörigkeitswünsche die Rezeption prägen. Handeln übersetzt die gewonnene Orientierung in eine unter den gegebenen Bedingungen verantwortbare Konsequenz.

8.2 Mathematische Weiterentwicklung: Geltungsräume und Unsicherheit

Die spätere mathematische Weiterentwicklung präzisiert jene Bereiche, die eine metaphorische Darstellung nur begrenzt erfassen kann. Sie modelliert Orientierung als getypten, dynamischen Zustand und hält Wissen, Grundvertrauen, konkrete Lebenssituation, Sinnhorizont, normative Bindungen sowie Kompetenzen und Ressourcen analytisch auseinander. Hinzu treten partielle Beobachtbarkeit, probabilistische Wissensaktualisierung, Kausalität, harte normative Grenzen, Mehrkriterienentscheidungen, Robustheit, Reversibilität und systemische Rückkopplung.19

Für KI-bezogene Entscheidungen erweist sich diese Architektur als hilfreich. Technische Leistungswerte werden dadurch daran gehindert, unbemerkt zu umfassenden Werturteilen anzuwachsen. Ein Modell kann bei der Vorhersagegenauigkeit überzeugen und gleichzeitig erhebliche Defizite bei Transparenz, Datenschutz, Fairness oder menschlicher Kontrolle aufweisen. Verantwortliche Entscheidungspraxis verlangt, diese Kriterien sichtbar zu machen, ihre Konflikte offenzulegen und Gewichtungen nachvollziehbar zu begründen.

---

18 Rieser, Der Weisheitskompass 2.0, insbesondere die Darstellung von Erkennen–Deuten, Fühlen–Handeln, Integration und Reifedimension: https://assets.sta.io/site_media/u/files/2026/01/31/Weisheitskompass_2.0_Rev_ygOMHc9.pdf

19 Rieser (2026), Verantwortliches Handeln unter Unsicherheit, unveröffentlichtes Manuskript; die Erweiterung verbindet getypte Geltungsräume, partielle Beobachtbarkeit, Bayes'sche Aktualisierung, normative Grenzen, Mehrkriterienentscheidungen, Robustheit, Informationswert, Reversibilität und Rückkopplung.

9. Erkennen und Deuten: epistemische Selbstständigkeit im Umgang mit KI

Erkennen beginnt mit der nüchternen Klärung dessen, was eine Aussage tatsächlich trägt. Bei einer KI-Antwort betrifft dies zunächst Quelle, Datenbasis, Aktualität, Unsicherheit und mögliche Fehler. Deuten erschließt anschließend die Bedeutung der Information im jeweiligen Zusammenhang. Diese Reihenfolge wirkt unspektakulär, schützt jedoch vor einem verbreiteten Kurzschluss: Eine statistisch wahrscheinliche Aussage kann im individuellen Fall unpassend bleiben, und eine elegante Formulierung kann eine fragwürdige Prämisse sprachlich überzeugend verbergen.

Epistemische Kompetenz bündelt mehrere Fähigkeiten: Quellen prüfen, probabilistische Aussagen erkennen, Gegenbelege suchen, Beobachtung von Kausalität unterscheiden und das eigene Urteil bei besserer Evidenz revidieren. Der mathematische Weisheitskompass beschreibt diese Haltung als revisionsoffene Wissensaktualisierung. Bayes'sches Denken fungiert dabei weniger als verpflichtendes Rechenverfahren denn als intellektuelle Disziplin: Überzeugungen erhalten vorläufige Plausibilität und müssen sich durch neue Evidenz verändern lassen.

Für generative KI folgt eine praktische Regel. Mit wachsender Tragweite einer Entscheidung steigen die Anforderungen an die Prüfung. Bei medizinischen, rechtlichen, finanziellen oder politischen Fragen reicht die sprachliche Plausibilität einer einzelnen Antwort keinesfalls aus; erforderlich werden unabhängige Quellen, fachliche Zuständigkeit und die sorgfältige Prüfung der konkreten Situation. KI kann die Vorbereitung solcher Entscheidungen verbessern. Die Verantwortung für ihre Folgen benötigt dennoch einen eindeutig bestimmbaren menschlichen oder institutionellen Träger.

10. Fühlen und Handeln: emotionale Resonanz, Persuasion und Verantwortung

Digitale Kommunikation trifft auf eine Aufmerksamkeit, die emotional selektiv arbeitet. Empörung, Angst, Hoffnung, Stolz und Zugehörigkeit beeinflussen, welche Inhalte erinnert, geteilt oder weiterverfolgt werden. Empfehlungssysteme können solche Reaktionen indirekt verstärken, sofern Engagement als Optimierungsziel dient. Die Analyse muss daher weniger nach einer vermeintlichen „Bosheit" des Algorithmus fragen als nach seiner Zielfunktion: Systeme lernen jene Muster zu bevorzugen, die das vorgegebene Erfolgskriterium steigern.

Der Weisheitskompass behandelt Gefühle als relevante Informationen über die eigene Situation. Eine starke emotionale Reaktion verdient deshalb Wahrnehmung und Reflexion, erhält dadurch jedoch keinen Wahrheitsstatus. Wer Intensität und Wahrheitsgehalt auseinanderhält, gewinnt Distanz zu Inhalten, die vor allem über Empörung oder Angst überzeugen. Diese Distanz verlangt keine emotionale Kälte; sie schafft vielmehr den Spielraum, in dem Gefühle als Orientierungsdaten ernst genommen und zugleich kritisch eingeordnet werden können.

Handlungskompetenz bewährt sich in diesem Zwischenraum. Schon eine kurze Verzögerung vor dem Teilen, Kaufen, Wählen oder Befolgen einer Empfehlung kann epistemischen Wert erzeugen. Je folgenreicher die Entscheidung, desto anspruchsvoller sollte die Prüfung ausfallen: alternative Quellen, sachkundige Personen, Risiken, Reversibilität und mögliche Interessenkonflikte gehören dann ausdrücklich in die Abwägung.

11. Integration und Reifung unter dem Risiko des Automation Bias

Reifung beschreibt im Weisheitskompass die wachsende Fähigkeit, Rückmeldungen aufzunehmen, unterschiedliche Perspektiven zusammenzuführen und die eigene Position zu korrigieren. Im Umgang mit KI erhält diese Lernfähigkeit unmittelbare praktische Bedeutung. Hochwertig formulierte Empfehlungen können einen Automation Bias fördern: Maschinelle Ausgaben erhalten dann ein Gewicht, das ihre tatsächliche Verlässlichkeit übersteigt, während eigene Hinweise auf Fehler oder Unangemessenheit zu wenig Beachtung finden.

Der europäische AI Act greift dieses Risiko für Hochrisiko-Systeme ausdrücklich auf. Artikel 14 verlangt menschliche Aufsicht und berücksichtigt die Gefahr einer übermäßigen oder fehlgeleiteten Orientierung an automatisierten Ausgaben. Aufsichtspersonen sollen Fähigkeiten und Grenzen des Systems verstehen, Ergebnisse sachgerecht interpretieren und Ausgaben erforderlichenfalls ignorieren, übersteuern oder rückgängig machen können. Rechtlich formulierte Aufsicht erhält damit einen erkenntnistheoretischen Kern: Wer Verantwortung trägt, benötigt reale Urteilsmacht.20

12. Orientierungs- und Handlungskompetenz als Bildungsauftrag

Orientierungskompetenz bezeichnet die Fähigkeit, in komplexen Informationslagen relevante Aussagen zu erkennen, ihre Geltungsart zu bestimmen, Quellen und Unsicherheiten einzuschätzen, emotionale Wirkungen wahrzunehmen und unterschiedliche Werte in einen nachvollziehbaren Zusammenhang zu bringen. Handlungskompetenz schließt daran an: Aus begründeter Orientierung entsteht ein konkreter Schritt, dessen Folgen beobachtet und bei Bedarf korrigiert werden.

Beide Kompetenzen bedingen einander. Eine ausschließlich prüfende Orientierung kann in endloser Analyse erstarren; ungeprüftes Handeln verwechselt leicht Geschwindigkeit mit Klugheit. Der Weisheitskompass verbindet deshalb Wahrnehmen, Verstehen, Prüfen, Deuten, Urteilen, Handeln und Revision in einem rekursiven Prozess. Die Schritte bilden keine starre Abfolge. Neue Informationen können frühere Deutungen verändern; Handlungsfolgen können neue Fragen eröffnen; Revision kann einen erneuten Prüfzyklus auslösen.

Für Bildung im KI-Zeitalter folgt eine deutliche Schwerpunktverschiebung. Sachwissen behält seinen Wert, denn Urteilskraft ohne Gegenstandskenntnis bleibt oberflächlich. Hinzutreten müssen Metakompetenzen: Fragen präzisieren, Quellen vergleichen, Unsicherheit aushalten, Interessen erkennen, Gegenmodelle prüfen, technische Ausgaben sinnvoll begrenzen und Verantwortung zuordnen. Je leistungsfähiger die Systeme werden, desto wichtiger wird die Fähigkeit, ihre Unterstützung zu nutzen, ohne die eigene Erkenntnis- und Urteilspraxis verkümmern zu lassen.

13. Transparenz, menschliche Aufsicht und institutionelle Verantwortung

Individuelle Kompetenz allein kann die Risiken algorithmischer Informationsumgebungen nicht auffangen. Technische und institutionelle Strukturen bestimmen, welche Einflussmöglichkeiten einzelnen Nutzerinnen und Nutzern überhaupt offenstehen. Internationale Leitlinien und europäische Regulierung setzen deshalb auf Transparenz, Menschenrechte, Aufsicht und Rechenschaft. Die UNESCO-Empfehlung zur Ethik

20 Verordnung (EU) 2024/1689 (AI Act), Art. 14 „Human oversight“: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

künstlicher Intelligenz stellt Menschenwürde und Menschenrechte in den Mittelpunkt; die OECD-Prinzipien betonen menschzentrierte Werte, Fairness, Transparenz und Erklärbarkeit.2122

In der Europäischen Union gewinnen die Transparenzpflichten des AI Act seit dem 2. August 2026 zusätzliche praktische Bedeutung. Die Kommission erläutert für bestimmte interaktive und generative Systeme Informations- und Kennzeichnungspflichten. Bei veröffentlichten KI-Texten zu Angelegenheiten von öffentlichem Interesse kommt darüber hinaus der menschlichen Überprüfung und redaktionellen Verantwortlichkeit besonderes Gewicht zu. Regulierung behandelt Transparenz damit zunehmend als Voraussetzung überprüfbarer Verantwortung und weniger als bloße technische Zusatzinformation.23

Der Weisheitskompass ergänzt die institutionelle Ebene um eine handlungspraktische Perspektive. Transparenz entfaltet nur dann Wirkung, wenn Menschen die bereitgestellten Informationen verstehen; Aufsicht bleibt leer, wenn keine reale Übersteuerungsmöglichkeit besteht; Erklärungen verlieren ihren Wert, wenn sie technisches Vokabular lediglich reproduzieren. Verantwortliche Governance verbindet daher Systemgestaltung, organisatorische Zuständigkeit und menschliche Kompetenz.

14. Hararis Warnung im Prüfverfahren des Weisheitskompasses

14.1 Wahrnehmen: Welche Veränderungen lassen sich beobachten?

Beobachtbar sind die rasche Verbreitung generativer KI, die Fähigkeit zu überzeugender Sprache, zunehmende Personalisierung und die zentrale Rolle algorithmischer Feeds in digitalen Öffentlichkeiten. Experimente weisen konkrete Persuasionseffekte und in einzelnen Plattformkontexten politische Einstellungsverschiebungen nach. Diese Befunde geben hinreichenden Anlass, mögliche Folgen frühzeitig und differenziert zu prüfen.

14.2 Verstehen: Welche Erklärungen konkurrieren?

Mehrere Erklärungen bleiben zugleich plausibel. Algorithmische Systeme können vorhandene Präferenzen spiegeln, sie verstärken, neue Inhalte erschließen oder Entscheidungen in bestimmte Richtungen lenken. Nutzerinnen und Nutzer wählen Plattformen teilweise selbst aus und werden zugleich von deren Architektur beeinflusst. Wirtschaftliche Anreize, soziale Identität, politische Akteure und individuelle Dispositionen greifen ineinander. Eine tragfähige Analyse muss diese Wechselwirkungen berücksichtigen, anstatt einen einzigen Kausalpfad zu unterstellen.

14.3 Prüfen: Welche Aussagen tragen empirische Evidenz?

Die Forschung trägt mehrere eng begrenzte Aussagen zuverlässig: Große Sprachmodelle können persuasive Texte erzeugen; Personalisierung kann die Überzeugungswirkung erhöhen; Feed-Algorithmen verändern nachweislich Exposition und teilweise politische Einstellungen. Hararis weiterreichende Prognosen über einen umfassenden Autoritätsverlust des Menschen überschreiten den gegenwärtigen empirischen

21 UNESCO, Recommendation on the Ethics of Artificial Intelligence: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

22 OECD.AI, Human-centred values and fairness: https://oecd.ai/en/dashboards/ai-principles/P6; Transparency and explainability: https://oecd.ai/en/dashboards/ai-principles/P7

23 Europäische Kommission, Guidelines on transparency obligations for providers and deployers of certain AI systems, Stand Juli/August 2026: https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations; FAQ zu Art. 50: https://digital-strategy.ec.europa.eu/en/faqs/transparency-obligations-under-article-50-ai-act

Nachweis. Ihr Erkenntniswert liegt vor allem darin, strukturelle Risiken sichtbar zu machen und Forschungsfragen zu schärfen.

14.4 Deuten: Welche Bedeutung erhält technische Assistenz?

Technische Assistenz kann als Entlastung, Erweiterung menschlicher Fähigkeiten, Statussymbol, Autorität oder Bedrohung erlebt werden. Solche Deutungen prägen den Umgang mit Systemen erheblich. Wer KI als nahezu unfehlbaren Experten auffasst, wird Empfehlungen anders gewichten als jemand, der sie als probabilistisches Werkzeug versteht. Reife Orientierung hält Leistungsfähigkeit, Grenzen und Kontext zugleich im Blick.

14.5 Urteilen: Welche Werte und Grenzen tragen die Entscheidung?

Autonomie, Menschenwürde, Wahrheitssuche, Sicherheit, Fairness, Privatheit und demokratische Verantwortlichkeit bilden eigenständige Bewertungskriterien. Effizienz kann mehrere dieser Güter fördern, erhält daraus jedoch keine automatische Vorrangstellung. Wo Anwendungen Zielkonflikte erzeugen, müssen Gewichtungen sichtbar und begründet werden; bei grundlegenden Schutzgütern setzen normative Grenzen dem Optimierungsdenken vorausgehende Schranken.

14.6 Handeln: Welche Schritte bleiben unter Unsicherheit tragfähig?

Robuste Handlungen bewähren sich unter mehreren plausiblen Zukunftsbildern. Dazu zählen transparente Kennzeichnung, unabhängige Überprüfung bei hohen Risiken, echte Übersteuerungsmöglichkeiten, begrenzte Pilotanwendungen, dokumentierte Verantwortlichkeiten sowie Bildung in Quellen- und Unsicherheitskompetenz. Solche Maßnahmen behalten ihren praktischen Wert auch dann, wenn Hararis dramatischste Szenarien ausbleiben.

14.7 Revision: Welche Rückmeldungen verändern das Modell?

Technik, Forschung und Regulierung entwickeln sich mit hoher Geschwindigkeit. Ein verantwortliches Modell muss deshalb neue Evidenz aufnehmen und seine Gewichtungen anpassen. Zeigen künftige Studien geringe gesellschaftliche Effekte, sind Risiken neu zu bewerten; weisen sie stärkere Manipulations- oder Abhängigkeitseffekte nach, müssen Schutzmechanismen entsprechend wachsen. Revision gehört damit zum Kern vernünftiger Steuerung.

15. Gesamturteil: Hararis Warnung zwischen Plausibilität und empirischer Tragfähigkeit

Hararis Warnung entfaltet ihren stärksten wissenschaftlichen Wert dort, wo sie auf Strukturprobleme aufmerksam macht: Konzentration von Informationsmacht, Personalisierung, asymmetrische Transparenz, persuasive Systeme und die Möglichkeit einer schleichenden Delegation menschlicher Urteilsarbeit. Die empirische Forschung bestätigt einzelne Mechanismen dieser Diagnose und setzt zugleich klare Grenzen für ihre Verallgemeinerung. Zwischen der nachweisbaren Wirkung einzelner Systeme und einer umfassenden historischen Prognose bleibt ein erheblicher Abstand.

Der Weisheitskompass verlagert die Debatte von der spektakulären Zukunftsprognose auf die Qualität gegenwärtiger Entscheidungen. Maßgeblich wird, ob Menschen und Institutionen Geltungsarten auseinanderhalten, Unsicherheit sichtbar machen, Gegenmodelle prüfen, normative Grenzen wahren, Rückkopplungen beobachten und Revision ermöglichen. Unter solchen Bedingungen kann KI als

leistungsfähiges Werkzeug in menschliche Entscheidungsprozesse eingebunden werden, ohne deren Verantwortungskern zu verdecken.

Daraus entsteht eine anspruchsvolle Form von Technikoffenheit. Sie würdigt reale Chancen und hält Schutzvorkehrungen zugleich für vernünftig. Pauschales Misstrauen erzeugt ebenso wenig Kompetenz wie ungeprüftes Vertrauen. Orientierung erwächst aus begründeter Unterscheidung, kritischer Prüfung und der Bereitschaft, Verantwortung auch dort sichtbar zu tragen, wo technische Systeme maßgeblich an einer Entscheidung mitwirken.

16. Schluss: Urteilshoheit als Aufgabe technischer Zivilisation

Die Herausforderung des KI-Zeitalters lässt sich kaum auf die Frage reduzieren, ob Maschinen eines Tages „intelligenter“ als Menschen werden. Bereits heute übertreffen technische Systeme menschliche Fähigkeiten in zahlreichen eng umrissenen Aufgaben. Gesellschaftlich drängender erscheint daher die Frage nach dem Ort menschlicher Urteilshoheit: Wer bestimmt Ziele, prüft Gründe, wägt Werte und Folgen ab und übernimmt Verantwortung für Entscheidungen, die unter technischer Mitwirkung zustande kommen?

Hararis Arbeiten machen die enge Verbindung von Informationsordnung und Macht sichtbar. Der Weisheitskompass ergänzt diese Diagnose um eine Theorie verantwortlichen Handelns unter Unsicherheit. Seine Logik zielt auf kompetente Kooperation mit technischen Systemen: Sie können rechnen, suchen, strukturieren, simulieren und Vorschläge entwickeln; Menschen und Institutionen bestimmen Ziele, Grenzen und Verantwortlichkeiten und beurteilen die Folgen ihres Handelns. Eine technikfreie Gegenwelt wird dafür weder vorausgesetzt noch angestrebt.

Weisheit bemisst sich unter solchen Bedingungen weniger am Umfang gespeicherten Wissens als an der Qualität des Umgangs damit. Sie verbindet Erkenntnis mit Deutung, emotionale Selbstwahrnehmung mit verantwortlichem Handeln und technische Möglichkeiten mit normativen Grenzen. Besonders deutlich zeigt sie sich in der Bereitschaft zur Revision: Wer bessere Gründe erkennt, kann seine Position verändern und gerade darin intellektuelle Verlässlichkeit beweisen.

Hararis Warnung lässt sich in eine konstruktive Leitfrage überführen: Wie lassen sich Informationssysteme und Bildungsprozesse so gestalten, dass technische Intelligenz menschliche Orientierungs- und Handlungskompetenz erweitert und deren schleichende Erosion verhindert? Der Weisheitskompass bietet keine fertige Zukunftsformel. Er stellt eine Architektur bereit, in der verschiedene Erkenntnisarten, Werte, Unsicherheiten und Handlungsmöglichkeiten geordnet, geprüft und in verantwortbare Entscheidungen überführt werden können.

Anhang A: Mathematisches Erklärmodell der Mensch-KI-Orientierungsdynamik

A.1 Zweck und wissenschaftlicher Status

Das folgende Modell übersetzt die im Essay beschriebene Wechselwirkung zwischen menschlicher Orientierung und algorithmisch geprägter Informationsumgebung in eine formale Sprache. Es versteht sich als eigenständige heuristische Weiterentwicklung im Rahmen des Weisheitskompasses. Die Formeln dienen der begrifflichen Disziplinierung und machen Abhängigkeiten sichtbar. Ohne empirische Operationalisierung, Reliabilitätsprüfung und Validierung haben sie keinen Status eines psychometrischen Messinstruments oder eines Prognosemodells.24

A.2 Menschlicher Orientierungszustand

Der menschliche Orientierungszustand wird zum Zeitpunkt t als mehrdimensionaler, getypter Zustand dargestellt:

$$X_t = (G_t, W_t, L_t, T_t, N_t, C_t)$$

G_t bezeichnet Grundvertrauen und grundlegende Überzeugungen, W_t den empirischen Wissensstand, L_t die konkrete Lebens- und Handlungssituation, T_t den Sinn- beziehungsweise Transzendenzhorizont, N_t normative Bindungen und C_t Kompetenzen sowie verfügbare Ressourcen. Eine Addition dieser Komponenten würde ihre unterschiedlichen Geltungsfunktionen verwischen; praktische Verbindung entsteht erst durch begründete Übergänge zwischen ihnen.

A.3 Algorithmische Informationsumgebung

Dem menschlichen Zustand wird eine algorithmisch geprägte Informationsumgebung gegenübergestellt:

$$A_t = (I_t, P_t, \Omega_t, E_t, R_t)$$

I_t erfasst das verfügbare Informationsangebot, P_t den Grad der Personalisierung, Ω_t die Opazität der Auswahl, E_t das emotionale Aktivierungspotenzial und R_t Wiederholung beziehungsweise algorithmische Verstärkung. Zwei Personen können somit theoretisch denselben Informationsraum betreten und praktisch sehr unterschiedliche Ausschnitte erleben.

A.4 Wahrgenommene Information

$$Y_t = \Phi(X_t, A_t) + \varepsilon_t$$

Y_t bezeichnet den tatsächlich wahrgenommenen Informationsausschnitt. Die Auswahlfunktion Φ hängt sowohl vom bisherigen Zustand der Person als auch von der technischen Umgebung ab; ε_t bündelt Zufall, nicht beobachtete Einflüsse und Modellfehler. Die Gleichung macht eine zentrale Differenz sichtbar: Verfügbares Wissen und tatsächlich wahrgenommene Information können in personalisierten Umgebungen systematisch auseinanderfallen.

24 Vgl. zur Selbstbegrenzung des mathematischen Ansatzes Rieser (2026), Verantwortliches Handeln unter Unsicherheit, Abschnitt „Wissenschaftlicher Status des Modells“: mathematische Form erhöht begriffliche Präzision, ersetzt jedoch keine empirische Operationalisierung und Validierung.

A.5 Entscheidung und Rückkopplung

$$a_t = \delta(X_t, Y_t)$$

$$X_{t+1} = F_x(X_t, Y_t, a_t, r_t, \xi_t)$$

$$A_{t+1} = F_a(A_t, \Psi(a_t))$$

a_t bezeichnet die Handlung, r_t die beobachteten Folgen und ξ_t nicht kontrollierbare Einflüsse. $\Psi(a_t)$ repräsentiert jene Verhaltensdaten, die aus einer Handlung hervorgehen und das technische System erneut informieren. In Alltagssprache beschreibt der Regelkreis eine fortlaufende Wechselwirkung: Verhalten erzeugt Daten, Daten verändern Auswahl, Auswahl prägt Wahrnehmung, und Wahrnehmung beeinflusst künftiges Verhalten.

A.6 Heuristischer algorithmischer Orientierungsdruck

Zur anschaulichen Verdichtung lässt sich ein heuristischer Orientierungsdruck definieren:

$$\Pi_t = P_t \cdot \Omega_t \cdot E_t \cdot R_t$$

Ein hoher Wert Π_t kennzeichnet eine Umgebung, in der Inhalte stark personalisiert, für Nutzer wenig durchschaubar, emotional aktivierend und häufig verstärkt werden. Die Multiplikation beruht auf keiner empirisch festgelegten Skala; sie veranschaulicht lediglich, wie mehrere Einflussbedingungen gemeinsam wirksam werden können.

A.7 Orientierungskompetenz

Der algorithmischen Einflussstruktur wird eine ebenfalls heuristische Kompetenzgröße K_t gegenübergestellt:

$$K_t = \Sigma_i \, w_i \, k_{i,t} \,, \quad \text{mit} \quad \Sigma_i \, w_i = 1$$

Die Teilkompetenzen $k_{i,t}$ können Quellenprüfung, Unsicherheitskompetenz, die Trennung von Tatsachen und Werturteilen, emotionale Selbstreflexion, Gegenperspektivenprüfung und Revisionsfähigkeit repräsentieren. Die Gewichte w_i legen offen, dass jede Zusammenfassung eine begründungsbedürftige Gewichtung voraussetzt.

A.8 Heuristischer Index der Orientierungsautonomie

$$S_t = K_t / (K_t + \Pi_t + \varepsilon)$$

ε verhindert eine Division durch null. S_t dient ausschließlich der Veranschaulichung: Je stärker die menschlichen Kompetenzen gegenüber dem algorithmischen Orientierungsdruck ausgeprägt sind, desto größer bleibt der Spielraum reflektierter Eigensteuerung. Eine psychometrische Deutung wäre ohne die empirische Entwicklung und Validierung eines Messinstruments methodisch unzulässig.

A.9 Bayes'sche Wissensaktualisierung

$$P(H \mid Y_t) = P(Y_t \mid H) \cdot P(H) / P(Y_t)$$

Die Bayes'sche Form erinnert daran, dass neue Evidenz die Plausibilität empirischer Hypothesen verändern soll. Für die Praxis genügt häufig die qualitative Frage, welche Beobachtung die eigene Annahme schwächen würde. Ein Informationssystem, das überwiegend bestätigende Inhalte liefert, kann subjektive Gewissheit erhöhen, obwohl die tatsächliche Evidenzlage offen bleibt.

A.10 Mehrkriterienentscheidung und normative Grenzen

Verantwortliche Entscheidungen lassen sich als Kriterienvektor darstellen:

$$V(a) = [\text{Evidenz, Autonomie, Fairness, Sicherheit, Privatheit, soziale Folgen, Reversibilität}]^T$$

Vor jeder Optimierung gelten normative Mindestbedingungen. Optionen, die grundlegende Rechte, Sicherheit oder andere verbindliche Schutzgrenzen verletzen, scheiden aus dem zulässigen Handlungsraum aus. Unter den verbleibenden Alternativen können Zielkonflikte transparent verglichen werden. Technische Effizienz erhält damit ihren sachgerechten Platz als eines von mehreren Kriterien.

A.11 Robustheit und Reversibilität

Unter Modellunsicherheit lässt sich ein robustes Entscheidungsprinzip schematisch als Minimax-Regret formulieren:

$$a^* = \arg\min_a \max_h \text{Regret}(a, h)$$

h bezeichnet verschiedene plausible Zukunftsmodelle. Die gewählte Handlung begrenzt den schlimmsten vermeidbaren Nachteil für den Fall, dass sich die bevorzugte Annahme später als falsch erweist. Ergänzend wird Reversibilität $\rho(a) \in [0,1]$ berücksichtigt. Wo keine Dringlichkeit besteht und Unsicherheit hoch bleibt, gewinnen beobachtbare und korrigierbare Schritte häufig einen höheren praktischen Wert als schwer rückgängig zu machende Festlegungen.

A.12 Informationswert

Neben unmittelbarem Nutzen kann eine Entscheidung auch Lernwert erzeugen. Schematisch:

$$J(a) = U(a) + \lambda \cdot IG(a) + \mu \cdot \rho(a) - \nu \cdot Risk(a)$$

U(a) bezeichnet den erwarteten Nutzen im mehrkriteriellen Sinn, IG(a) den Informationsgewinn, ρ(a) die Reversibilität und Risk(a) relevante Risiken. λ, μ und ν markieren offene Gewichtungen. Das Modell zwingt dazu, diese Gewichtungen sichtbar zu machen; ihre moralische Begründung kann es selbst nicht liefern.

A.13 Gesamtmodell

$$X_t \rightarrow Y_t \rightarrow a_t \rightarrow \Psi(a_t) \rightarrow A_{t+1} \rightarrow Y_{t+1} \rightarrow X_{t+1}$$

Der mathematische Kern beschreibt eine gekoppelte Dynamik. Mensch und Informationssystem verändern wechselseitig die Bedingungen künftiger Entscheidungen. Eine Momentaufnahme der vermeintlich „richtigen Information“ reicht deshalb häufig nicht aus. Verantwortliche Orientierung beobachtet ebenso, welche Gewohnheiten, Abhängigkeiten, Auswahlmuster und sozialen Rückkopplungen aus wiederholter Nutzung hervorgehen.

A.14 Verdichtete Aussage

$$\text{Mehr Information} \neq \text{mehr Wissen} \neq \text{mehr Weisheit}$$

Information verlangt Prüfung, Wissen verlangt Begründung, verantwortliches Handeln verlangt Abwägung und Rückkopplung. Der formale Mehrwert des Modells besteht in der sichtbaren Trennung dieser Ebenen.

Anhang B: Variablen und Begriffe des mathematischen Erklärmodells

Die folgende Übersicht übersetzt die im mathematischen Erklärmodell verwendeten Variablen, Zeichen und Fachbegriffe in eine allgemein verständliche Sprache. Sie soll das Lesen der Formeln erleichtern, ohne ihnen eine höhere empirische Genauigkeit zuzuschreiben, als das Modell tatsächlich beansprucht. Die Symbolwahl dient der formalen Übersicht; mehrere Zeichen – etwa Π für Orientierungsdruck oder Ω für Opazität – sind Arbeitsnotationen dieses Modells und keine allgemein verbindlichen Standards der Mathematik oder KI-Forschung.

Zeichen / Begriff	Bedeutung im Modell	Verständlich erklärt
Zeit und Grundstruktur		
t	Zeitpunkt der Betrachtung	Das kleine t bedeutet: Wir betrachten die Situation zu einem bestimmten Zeitpunkt. Dadurch kann das Modell Veränderungen über die Zeit beschreiben.
t + 1	Nächster Zeitpunkt	Damit ist der Zustand nach einem weiteren Schritt gemeint – etwa nachdem eine Information gesehen, eine Entscheidung getroffen und eine Rückmeldung erhalten wurde.
X_t	Menschlicher Orientierungszustand	X_t fasst jene menschlichen Bedingungen zusammen, die für eine Entscheidung gerade bedeutsam sind: Grundüberzeugungen, Wissen, Lebenssituation, Werte, Sinnfragen und verfügbare Fähigkeiten.
G_t	Grundvertrauen und grundlegende Überzeugungen	Dazu gehören tiefere Annahmen darüber, wem oder was man vertraut und welche Grundüberzeugungen den Blick auf eine Situation prägen.
W_t	Empirischer Wissensstand	Gemeint ist das gegenwärtig verfügbare, sachlich begründete Wissen. Es kann wachsen, korrigiert werden oder sich bei neuer Evidenz verändern.
L_t	Konkrete Lebens- und Handlungssituation	Diese Variable erinnert daran, dass dieselbe Information je nach Person, Aufgabe, Zeitpunkt und Umfeld unterschiedliche praktische Bedeutung gewinnen kann.
T_t	Sinn- und Transzendenzhorizont	Hier werden Sinnfragen, weltanschauliche oder religiöse Deutungen erfasst. Sie gehören zur Orientierung, folgen jedoch anderen Prüfregeln als empirische Tatsachenbehauptungen.
N_t	Normative Bindungen	Damit sind Werte, Pflichten und moralische Grenzen gemeint – beispielsweise Würde, Fairness, Verantwortung oder der Schutz anderer Menschen.
C_t	Kompetenzen und Ressourcen	C_t beschreibt, was eine Person tatsächlich kann und worauf sie zurückgreifen kann: Wissen, Erfahrung, Zeit, Unterstützung, technische Fähigkeiten oder institutionelle Möglichkeiten.
Algorithmische Informationsumgebung		
A_t	Algorithmisch geprägte Informationsumgebung	A_t fasst Eigenschaften der digitalen Umgebung zusammen, in der Informationen ausgewählt, sortiert und präsentiert werden.
I_t	Verfügbares Informationsangebot	Das ist die Menge an Informationen, die prinzipiell erreichbar wäre. Sie ist meist erheblich größer als das, was eine Person tatsächlich sieht.
P_t	Grad der Personalisierung	P_t beschreibt, wie stark das System Inhalte auf eine bestimmte Person zuschneidet – etwa anhand früherer Klicks, Suchanfragen oder Interessen. P_t darf nicht mit der Wahrscheinlichkeitsnotation P(H) verwechselt werden.
Ω_t (Omega)	Opazität oder Undurchschaubarkeit	Ω_t steht dafür, wie schwer nachvollziehbar die Auswahl des Systems bleibt. Ein hoher Wert bedeutet im Modell: Die Gründe für die angezeigten Inhalte sind für Nutzerinnen und Nutzer wenig transparent.

Zeichen / Begriff	Bedeutung im Modell	Verständlich erklärt
E_t	Emotionales Aktivierungspotenzial	Damit wird bezeichnet, wie stark Inhalte voraussichtlich Aufmerksamkeit, Angst, Empörung, Hoffnung, Zugehörigkeit oder andere emotionale Reaktionen ansprechen.
R_t	Wiederholung und algorithmische Verstärkung	R_t erfasst, wie häufig oder wie nachdrücklich bestimmte Inhalte erneut sichtbar gemacht werden. Wiederholung kann Vertrautheit und subjektive Bedeutsamkeit erhöhen.
Wahrnehmung, Entscheidung und Rückkopplung		
Y_t	Tatsächlich wahrgenommener Informationsausschnitt	Y_t bezeichnet nicht das gesamte verfügbare Wissen, sondern genau jenen Ausschnitt, der die Person in der konkreten Situation tatsächlich erreicht.
Φ (Phi)	Auswahlfunktion	Φ ist ein Platzhalter für den komplexen Auswahlvorgang: Welche Inhalte werden aus der vorhandenen Informationsmenge gerade sichtbar? Dabei wirken sowohl Eigenschaften der Person als auch des technischen Systems mit.
ε_t (Epsilon)	Zufall und nicht erfasste Einflüsse	Kein Modell kann alle Ursachen kennen. ε_t erinnert an Zufälle, Messfehler und sonstige Einflüsse, die in den übrigen Variablen nicht ausdrücklich enthalten sind.
a_t	Handlung oder Entscheidung	a_t bezeichnet den konkreten Schritt, den eine Person zum Zeitpunkt t setzt – beispielsweise einen Inhalt anklicken, eine Empfehlung annehmen, widersprechen oder weitere Informationen suchen.
δ (Delta)	Entscheidungsfunktion	δ steht abstrakt für den Vorgang, durch den aus persönlichem Zustand und wahrgenommener Information eine Handlung hervorgeht. Die Formel behauptet nicht, diesen Vorgang vollständig berechnen zu können.
r_t	Beobachtete Folgen	r_t bezeichnet das, was nach einer Handlung als Ergebnis oder Rückmeldung wahrgenommen wird. Solche Folgen können das spätere Urteil verändern.
ξ_t (Xi)	Nicht kontrollierbare Einflüsse	ξ_t hält fest, dass Folgen auch von Umständen abhängen, die weder die Person noch das System kontrollieren können.
F_x	Veränderungsregel des menschlichen Zustands	F_x beschreibt formal, wie sich der Orientierungszustand nach neuen Informationen, Entscheidungen und Erfahrungen weiterentwickelt.
F_a	Veränderungsregel der algorithmischen Umgebung	F_a beschreibt, wie das technische System auf neue Nutzungsdaten reagiert und dadurch die künftige Informationsauswahl verändern kann.
$\Psi(a_t)$ (Psi)	Aus einer Handlung entstehende Nutzungsdaten	Ein Klick, eine Verweildauer, ein Like oder eine Suchanfrage erzeugt Daten. $\Psi(a_t)$ fasst solche Datenspuren zusammen, die das System anschließend zur weiteren Auswahl verwenden kann.
Rückkopplung / Regelkreis	Wechselseitige Veränderung	Menschen beeinflussen durch ihr Verhalten das System; das System verändert daraufhin die Informationsumgebung; diese neue Umgebung wirkt wieder auf Wahrnehmung und Entscheidungen zurück.
Heuristische Größen für Einfluss und Kompetenz		
Π_t (Pi)	Heuristischer algorithmischer Orientierungsdruck	Π_t verdichtet Personalisierung, Undurchschaubarkeit, emotionale Aktivierung und Wiederholung zu einer einzigen Anschauungsgröße. Sie dient dem Denken über Zusammenhänge, nicht einer bereits validierten Messung.
K_t	Heuristische Orientierungskompetenz	K_t fasst Fähigkeiten zusammen, die selbstständige Orientierung unterstützen: Quellen prüfen, Unsicherheit erkennen, Werte von Tatsachen unterscheiden, Emotionen reflektieren, Gegenpositionen prüfen und Urteile revidieren.
$k_{i,t}$	Einzelne Teilkompetenz	Das i steht für verschiedene Teilfähigkeiten. Eine Teilkompetenz kann etwa Quellenkritik oder Revisionsfähigkeit sein; t zeigt wiederum den betrachteten Zeitpunkt an.

Zeichen / Begriff	Bedeutung im Modell	Verständlich erklärt
w_i	Gewicht einer Teilkompetenz	Nicht jede Fähigkeit muss in jeder Entscheidung gleich wichtig sein. w_i gibt an, wie stark eine Teilkompetenz in die zusammenfassende Größe K_t eingeht.
Σ_i (Sigma)	Summenzeichen	Σ bedeutet: Mehrere Einzelwerte werden zusammengezählt. Im Modell werden die gewichteten Teilkompetenzen zu K_t zusammengefasst.
S_t	Heuristischer Index der Orientierungsautonomie	S_t veranschaulicht das Verhältnis zwischen menschlicher Orientierungskompetenz und algorithmischem Orientierungsdruck. Ein höherer Wert soll einen größeren Spielraum reflektierter Eigensteuerung ausdrücken; ein validierter psychologischer Messwert liegt damit nicht vor.
ε im Nenner	Kleine technische Schutzgröße	Dieses ε verhindert, dass in einem Sonderfall durch null geteilt wird. Es trägt keine eigenständige psychologische oder gesellschaftliche Bedeutung.
Wahrscheinlichkeit und Wissensaktualisierung		
H	Hypothese	H bezeichnet eine Annahme über die Wirklichkeit, die durch neue Informationen gestützt oder geschwächt werden kann.
P(H)	Ausgangsplausibilität einer Hypothese	P(H) steht dafür, wie plausibel eine Annahme vor der neuen Information eingeschätzt wird. Das P bedeutet hier Wahrscheinlichkeit und hat nichts mit der Personalisierungsvariable P_t zu tun.
$P(Y_t \mid H)$	Wahrscheinlichkeit der Beobachtung unter der Hypothese	Die Schreibweise fragt: Wie gut wäre die beobachtete Information Y_t zu erwarten, wenn die Hypothese H zuträfe?
$P(H \mid Y_t)$	Aktualisierte Plausibilität nach neuer Information	Hier wird gefragt: Wie plausibel erscheint H, nachdem Y_t berücksichtigt wurde? Diese Aktualisierung bildet den Kern des Bayes'schen Denkens.
Bayes'sche Aktualisierung	Lernen aus neuer Evidenz	Eine Überzeugung bleibt vorläufig und verändert sich, wenn neue, relevante Informationen hinzukommen. Praktisch bedeutet das: Gute Gründe dürfen die eigene Einschätzung stärken oder schwächen.
Mehrkriterienentscheidung, Robustheit und Lernwert		
V(a)	Kriterienvektor einer Handlung	V(a) hält mehrere Gesichtspunkte gleichzeitig fest – etwa Evidenz, Autonomie, Fairness, Sicherheit, Privatheit, soziale Folgen und Reversibilität. So wird eine komplexe Entscheidung nicht auf eine einzige Kennzahl verkürzt.
$[\dots]^T$	Vektorschreibweise	Die eckige Klammer fasst mehrere Werte zu einem geordneten Bündel zusammen. Das hochgestellte T bedeutet mathematisch eine Transposition und kennzeichnet hier lediglich die Darstellung als Spaltenvektor.
a	Mögliche Handlungsalternative	a steht allgemein für eine zur Wahl stehende Handlung. a_t bezeichnet die tatsächlich gewählte Handlung zu einem bestimmten Zeitpunkt.
h	Plausibles Zukunfts- oder Erklärungsmodell	h bezeichnet eine mögliche Annahme darüber, wie die Welt funktioniert oder wie sich eine Situation entwickeln könnte. Mehrere h erlauben, Entscheidungen unter verschiedenen Zukunftsbildern zu prüfen.
Regret(a,h)	Bedauerns- bzw. Verlustgröße	Regret beschreibt, wie viel schlechter eine gewählte Handlung im Nachhinein gegenüber der besten Alternative unter einem bestimmten Zukunftsmodell h ausgefallen wäre.
$\arg\min_a \max_h$	Suche nach einer robusten Handlung	Die Zeichen bedeuten vereinfacht: Wähle jene Handlung, bei der der schlimmste vermeidbare Nachteil über die betrachteten Zukunftsszenarien möglichst klein bleibt.
Minimax-Regret	Robustheitsprinzip	Dieses Entscheidungsprinzip bevorzugt eine Lösung, die sich auch dann noch vernünftig vertreten lässt, wenn die bevorzugte Annahme später nicht zutrifft.

Zeichen / Begriff	Bedeutung im Modell	Verständlich erklärt
ρ(a) (Rho)	Reversibilität einer Handlung	ρ(a) beschreibt, wie gut eine Entscheidung später korrigiert oder zurückgenommen werden kann. Werte nahe 1 stehen im Modell für hohe, Werte nahe 0 für geringe Reversibilität.
∈ [0,1]	Wertebereich zwischen 0 und 1	Die Schreibweise bedeutet: Die betreffende Größe liegt zwischen 0 und 1. Null und Eins markieren dabei die beiden Endpunkte einer vereinfachten Skala.
J(a)	Gesamtbewertung einer Handlung	J(a) verbindet im Modell erwarteten Nutzen, Lernwert, Reversibilität und Risiko. Die Formel macht sichtbar, welche Gesichtspunkte in eine Entscheidung einfließen sollen.
U(a)	Erwarteter Nutzen	U(a) steht für den erwarteten praktischen Wert einer Handlung. Gemeint ist kein bloßer Geldwert, sondern eine mehrkriterielle Bewertung.
IG(a)	Informationsgewinn	IG bedeutet Information Gain. Eine Handlung kann wertvoll sein, weil sie neue, entscheidungsrelevante Erkenntnisse liefert – selbst wenn sie noch keine endgültige Lösung darstellt.
Risk(a)	Relevantes Risiko	Risk(a) bündelt mögliche Schäden oder unerwünschte Folgen einer Handlung, soweit sie für die Entscheidung berücksichtigt werden sollen.
λ, μ, ν (Lambda, My, Ny)	Gewichtungsfaktoren	Diese griechischen Buchstaben geben an, wie stark Informationsgewinn, Reversibilität und Risiko in die Gesamtbewertung eingehen. Ihre Wahl verlangt eine begründete Entscheidung; die Mathematik liefert diese Wertung nicht von selbst.
Normative Mindestbedingungen	Grenzen vor jeder Optimierung	Bestimmte Handlungen können trotz hoher Effizienz ausscheiden, wenn sie grundlegende Rechte, Sicherheit, Würde oder andere verbindliche Schutzanforderungen verletzen.
Wissenschaftlicher Status des Modells		
Heuristisch	Denk- und Erklärhilfe	Eine heuristische Formel ordnet Zusammenhänge und macht Fragen sichtbar. Sie beansprucht noch nicht, die Wirklichkeit mit empirisch bestätigter Genauigkeit messen oder vorhersagen zu können.
Operationalisierung	Übersetzung eines Begriffs in messbare Merkmale	Bevor etwa „Orientierungskompetenz“ empirisch gemessen werden könnte, müsste genau festgelegt werden, welche beobachtbaren Aufgaben oder Fragen diese Fähigkeit zuverlässig erfassen.
Reliabilität	Zuverlässigkeit einer Messung	Ein reliables Messinstrument liefert unter vergleichbaren Bedingungen möglichst stabile und reproduzierbare Ergebnisse.
Validität	Gültigkeit einer Messung	Validität fragt, ob ein Verfahren tatsächlich das misst, was es zu messen vorgibt. Eine präzise Zahl kann unbrauchbar sein, wenn sie den falschen Gegenstand erfasst.
Psychometrisches Messinstrument	Empirisch geprüftes Verfahren zur Messung psychologischer Merkmale	Dafür wären definierte Skalen, Testaufgaben, Vergleichsdaten sowie Untersuchungen zu Reliabilität und Validität erforderlich. Das vorliegende mathematische Erklärmodell erfüllt diesen Anspruch ausdrücklich noch nicht.
Prognosemodell	Modell zur Vorhersage zukünftiger Ergebnisse	Ein Prognosemodell müsste anhand realer Daten zeigen, dass seine Vorhersagen außerhalb der Entwicklungsdaten ausreichend genau funktionieren. Auch diesen Anspruch erhebt das vorliegende Modell nicht.

Leseregel: Die Formeln sollen Beziehungen sichtbar machen. Wo eine Variable nicht empirisch operationalisiert und validiert wurde, darf aus ihrer mathematischen Schreibweise keine scheinbare Messgenauigkeit abgeleitet werden. Für das Verständnis genügt daher zunächst die Frage, welche Rolle ein Zeichen im Gedankengang übernimmt und welche Annahmen damit verbunden sind.

Anhang C: Sieben Praxisfragen für den verantwortlichen Umgang mit KI

Praxisfrage	Leitfrage für die Anwendung
1. Wahrnehmen	Welche Information liegt tatsächlich vor? Welche Quelle, welches System und welcher Kontext lassen sich erkennen? Welche körperlichen oder emotionalen Reaktionen treten auf?
2. Verstehen	Welche alternativen Erklärungen kommen in Betracht? Welche Interessen, Auswahlmechanismen oder sozialen Rückkopplungen könnten die Information geprägt haben?
3. Prüfen	Welche Aussage beansprucht empirische Geltung, welche interpretiert und welche wertet? Sind die Quellen auffindbar, aktuell und hinreichend unabhängig? Welche Gegenbelege wären denkbar?
4. Deuten	Welche Bedeutung erhält die Empfehlung im konkreten Zusammenhang? Wodurch gewinnt sie Überzeugungskraft, Bedrohlichkeit oder entlastende Wirkung? Welche Rolle spielen Vertrauen, Status und Gruppenzugehörigkeit?
5. Urteilen	Welche Werte, Pflichten und Schutzgrenzen berührt die Entscheidung? Wer trägt mögliche Folgen, und wo liegt die tatsächliche Verantwortung?
6. Handeln	Welcher nächste Schritt bleibt unter mehreren plausiblen Annahmen vertretbar? Lässt er sich begrenzen, beobachten und erforderlichenfalls korrigieren oder rückgängig machen?
7. Revision	Welche Folgen haben sich nach der Handlung gezeigt? Welche Annahmen wurden gestützt oder geschwächt? Welche Anpassungen am eigenen Urteil oder am Einsatz des Systems werden erforderlich?

Die sieben Fragen bilden kein starres Prüfschema. Reale Entscheidungssituationen führen durch neue Informationen häufig zu früheren Schritten zurück. Gerade diese Rückbewegung macht Revisionsfähigkeit praktisch sichtbar.

Anhang D: Erklärübersicht zentraler Fachbegriffe

Fachbegriff	Verständliche Erklärung
Algorithmischer Feed	Automatisch zusammengestellte Reihenfolge von Inhalten. Daten über Nutzung und vorgegebene Zielgrößen bestimmen mit, welche Beiträge sichtbar werden und in welcher Reihenfolge sie erscheinen.
Empfehlungssystem	Verfahren, das aus vielen möglichen Optionen jene auswählt, die für eine Person oder Gruppe als besonders passend eingeschätzt werden.

Personalisierung	Anpassung von Inhalten oder Empfehlungen an Merkmale, Präferenzen und bisheriges Verhalten einzelner Nutzerinnen und Nutzer.
Opazität	Begrenzte Durchschaubarkeit eines Systems. Betroffene können dann nur eingeschränkt nachvollziehen, weshalb eine bestimmte Empfehlung oder Auswahl zustande kam.
Persuasion	Überzeugende Kommunikation, die Einstellungen oder Entscheidungen beeinflussen kann. Der Begriff umfasst legitime Argumentation ebenso wie problematische Formen gezielter Einflussnahme.
Manipulation	Gezielte Einflussnahme, die selbstständige Urteilsbildung in problematischer Weise umgeht, verzerrt oder ausnutzt.
Automation Bias	Neigung, automatisierten Empfehlungen ein übermäßiges Gewicht zu geben und eigene Hinweise auf Fehler oder Unangemessenheit zu unterschätzen.
Autonomie	Fähigkeit und reale Möglichkeit, Entscheidungen aufgrund eigener reflektierter Gründe zu treffen und äußere Einflussnahmen kritisch zu verarbeiten.
Orientierungskompetenz	Fähigkeit, Informationen, Deutungen, Werte und Unsicherheiten so zu ordnen, dass ein begründetes Urteil entstehen kann.
Handlungskompetenz	Fähigkeit, aus begründeter Orientierung einen angemessenen Schritt abzuleiten, Verantwortung zu übernehmen und die Folgen der Entscheidung zu beobachten.
Geltungsart	Art des Anspruchs, den eine Aussage erhebt. Empirische Aussagen verlangen andere Prüfverfahren als Interpretationen, Werturteile oder Glaubensaussagen.
Evidenz	Gesamtheit der Belege, die eine Aussage stützen oder schwächen. Zur Bewertung gehören unter anderem Qualität, Unabhängigkeit, Messgenauigkeit und die Prüfung alternativer Erklärungen.
Kausalität	Begründete Ursache-Wirkungs-Beziehung. Gemeinsames Auftreten oder zeitliche Reihenfolge allein tragen noch keinen Kausalnachweis.
Bayes’sche Aktualisierung	Formale Denkregel, nach der neue Evidenz die Plausibilität einer Hypothese verändert. Praktisch fordert sie, Überzeugungen für Korrektur durch neue Daten offenzuhalten.
Mehrkriterienentscheidung	Entscheidungsverfahren, das mehrere berechtigte Ziele gleichzeitig berücksichtigt, etwa Effizienz, Autonomie, Sicherheit, Fairness und Reversibilität.
Robustheit	Eigenschaft einer Entscheidung, auch unter anderen plausiblen Annahmen noch vertretbar zu bleiben.

Reversibilität	Möglichkeit, einen Schritt später zu korrigieren oder zurückzunehmen. Unter hoher Unsicherheit kann diese Eigenschaft besonderen praktischen Wert gewinnen.
Informationswert	Nutzen einer Handlung, der daraus entsteht, dass sie neue Erkenntnisse erzeugt und spätere Entscheidungen verbessert.
Rückkopplung	Prozess, bei dem Folgen einer Handlung auf die Ausgangssituation zurückwirken. Bei Plattformen beeinflusst Nutzerverhalten beispielsweise die spätere Auswahl von Inhalten.
Menschliche Aufsicht	Organisierte Möglichkeit, ein KI-System zu beobachten, seine Grenzen zu verstehen und Ausgaben bei Bedarf zu korrigieren, zu übersteuern oder zu verwerfen.

Quellen und Literatur

Die nachstehenden Webadressen sind als direkte externe Hyperlinks hinterlegt und führen ohne Weiterleitung über ChatGPT oder OpenAI unmittelbar zu den jeweiligen Originalseiten beziehungsweise DOI-Zielen.

Harari, Yuval Noah: Homo Deus. A Brief History of Tomorrow. Offizielle Werkseite. https://www.ynharari.com/book/homo-deus/

Harari, Yuval Noah: 21 Lessons for the 21st Century. Offizielle Werkseite. https://www.ynharari.com/book/21-lessons-book/

Harari, Yuval Noah: Nexus. A Brief History of Information Networks from the Stone Age to AI. Offizielle Werkseite. https://www.ynharari.com/book/nexus/

Harari, Yuval Noah: Nexus – Notes / Quellenapparat, 2024. https://www.ynharari.com/wp-content/uploads/2024/11/Notes-NEXUS-Yuval-Noah-Harari-September-2024.pdf

SRF, Sternstunde Philosophie: Yuval Harari – Ein Historiker erzählt die Geschichte von morgen. https://www.srf.ch/play/tv/sternstunde-philosophie/video/yuval-harari-ein-historiker-erzaehlt-die-geschichte-von-morgen?urn=urn%3Asrf%3Avideo%3A8b1ea2c8-0be4-44cb-82bc-6c3d7631ceba

The Diary of a CEO: Yuval Noah Harari: They Are Lying About AI! The Trump Kamala Election Will Tear The Country Apart!, 5. September 2024. https://www.youtube.com/watch?v=78YN1e8UXdM

Bai, Hui; Voelkel, Jan G.; Muldowney, Shane; Eichstaedt, Johannes C.; Willer, Robb (2025): LLM-generated messages can persuade humans on policy issues. Nature Communications 16, 6037. DOI: https://doi.org/10.1038/s41467-025-61345-5

Salvi, Francesco; Horta Ribeiro, Manoel; Gallotti, Riccardo; West, Robert (2025): On the conversational persuasiveness of GPT-4. Nature Human Behaviour 9, 1645–1653. DOI: https://doi.org/10.1038/s41562-025-02194-6

Nyhan, Brendan et al. (2023): Like-minded sources on Facebook are prevalent but not polarizing. Nature 620, 137–144. DOI: https://doi.org/10.1038/s41586-023-06297-w

Gauthier, Germain; Hodler, Roland; Widmer, Philine; Zhuravskaya, Ekaterina (2026): The political effects of X's feed algorithm. Nature 652, 416–423. DOI: https://doi.org/10.1038/s41586-026-10098-2

Bonicalzi, Sofia; De Caro, Mario; Giovanola, Benedetta (2023): Artificial Intelligence and Autonomy: On the Ethical Dimension of Recommender Systems. Topoi 42, 819–832. DOI: https://doi.org/10.1007/s11245-023-09922-5

Milano, Silvia; Taddeo, Mariarosaria; Floridi, Luciano (2020): Recommender systems and their ethical challenges. AI & Society 35, 957–967. DOI: https://doi.org/10.1007/s00146-020-00950-y

Catena, Eleonora (2026): AI and human autonomy: a literature review. AI and Ethics 6, Article 126. DOI: https://doi.org/10.1007/s43681-025-00958-4

UNESCO: Recommendation on the Ethics of Artificial Intelligence. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

OECD.AI: Human-centred values and fairness – OECD AI Principle. https://oecd.ai/en/dashboards/ai-principles/P6

OECD.AI: Transparency and explainability – OECD AI Principle. https://oecd.ai/en/dashboards/ai-principles/P7

Europäische Union: Verordnung (EU) 2024/1689 – Artificial Intelligence Act, insbesondere Art. 13–14 und 50. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Europäische Kommission (2026): Guidelines on transparency obligations for providers and deployers of certain AI systems. https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations

Rieser, Norbert (2026): Der Weisheitskompass 2.0. Versuch einer existenziellen Orientierung im Horizont zeitgemäß reflektierten Glaubens. Öffentliche PDF-Fassung. https://assets.sta.io/site_media/u/files/2026/01/31/Weisheitskompass_2.0_Rev_ygOMHc9.pdf

Rieser, Norbert (2026): Verantwortliches Handeln unter Unsicherheit. Eine erkenntnistheoretisch, systemisch und mathematisch weiterentwickelte Handlungstheorie im Rahmen des Weisheitskompasses.

Auf Homepage veröffentlichtes Manuskript.

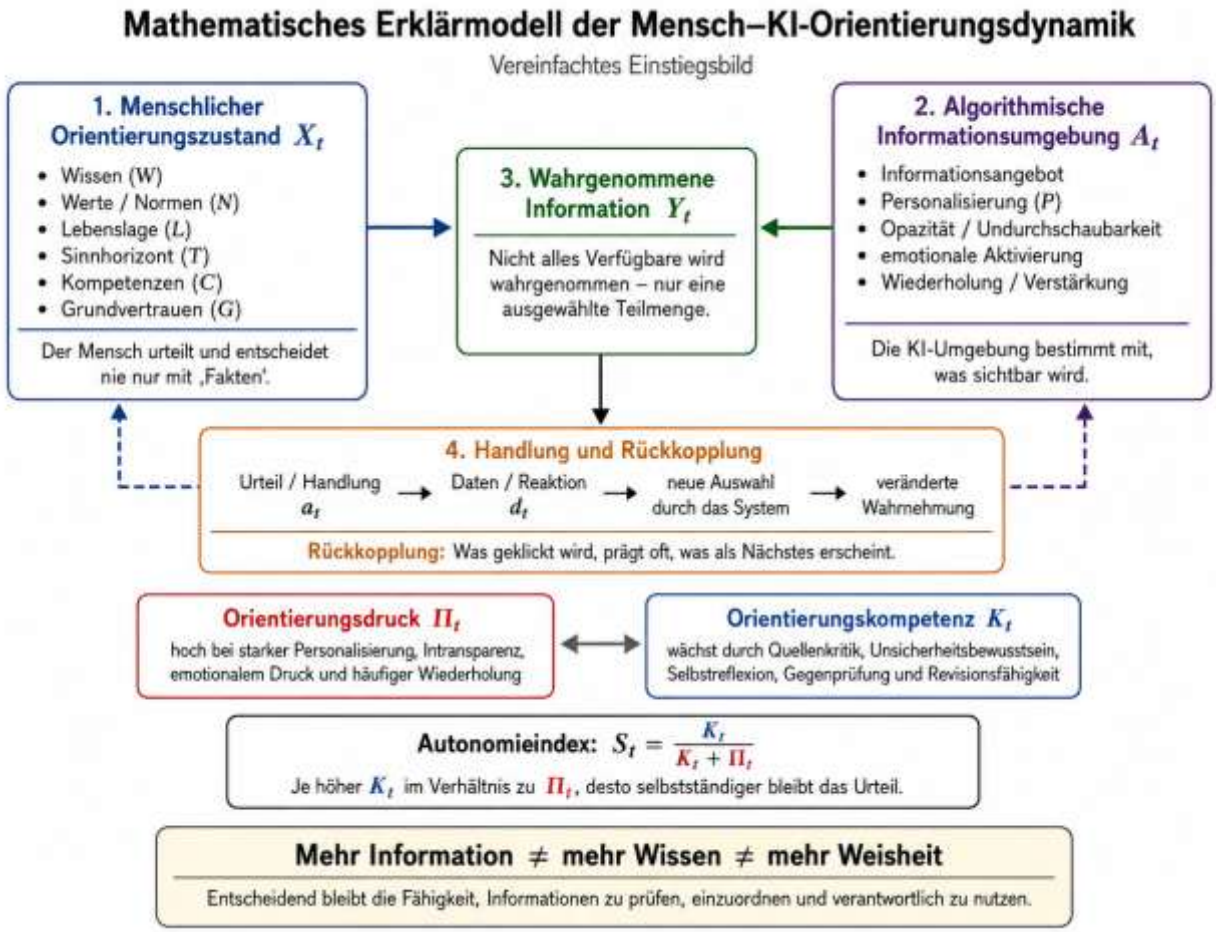
Rieser, Norbert (2026): Praxisanhang zum mathematisch weiterentwickelten Weisheitskompass. Verständliche Anwendungsbeispiele zu Erkenntnis, Entscheidung, Verantwortung und Revision.

Auf Homepage veröffentlichtes Manuskript.

Schlussbemerkung zur Nutzung

Dieser Essay dient der wissenschaftlichen Orientierung und der Weiterentwicklung von Orientierungs- und Handlungskompetenz im Umgang mit künstlicher Intelligenz. Für konkrete medizinische, rechtliche, finanzielle oder sicherheitskritische Einzelfälle kann er fachliche Beratung weder vorwegnehmen noch ersetzen. Wo Entscheidungen erhebliche Folgen für Gesundheit, Rechte, Vermögen oder andere geschützte Güter entfalten können, sollten einschlägig qualifizierte Fachpersonen einbezogen und die maßgeblichen Ausgangsdaten gemeinsam geprüft werden.

Die dynamische Entwicklung von KI-Systemen, Forschung und Regulierung verlangt fortlaufende Revision. Wissenschaftliche Befunde, technische Fähigkeiten und rechtliche Anforderungen verändern sich; dementsprechend müssen auch die hier vorgenommenen Bewertungen bei neuer Evidenz überprüft und gegebenenfalls angepasst werden.


Mathematisches Erklärmodell der Mensch–KI-Orientierungsdynamik
Vereinfachtes Einstiegsbild
1. Menschlicher Orientierungszustand X_t
Wissen (W)
Werte / Normen (N)
Lebenslage (L)
Sinnhorizont (T)
Kompetenzen (C)
Grundvertrauen (G)
Der Mensch urteilt und entscheidet nie nur mit ‚Fakten'.
3. Wahrgenommene Information Y_t
Nicht alles Verfügbare wird wahrgenommen – nur eine ausgewählte Teilmenge.
2. Algorithmische Informationsumgebung A_t
Informationsangebot
Personalisierung (P)
Opazität / Undurchschaubarkeit
emotionale Aktivierung
Wiederholung / Verstärkung
Die KI-Umgebung bestimmt mit, was sichtbar wird.
4. Handlung und Rückkopplung
Urteil / Handlung a_t
Daten / Reaktion d_t
neue Auswahl durch das System
veränderte Wahrnehmung
Rückkopplung: Was geklickt wird, prägt oft, was als Nächstes erscheint.
Orientierungsdruck Π_t
hoch bei starker Personalisierung, Intransparenz, emotionalem Druck und häufiger Wiederholung
Orientierungskompetenz K_t
wächst durch Quellenkritik, Unsicherheitsbewusstsein, Selbstreflexion, Gegenprüfung und Revisionsfähigkeit
Autonomieindex: $S_t = \frac{K_t}{K_t + \Pi_t}$
Je höher K_t im Verhältnis zu Π_t, desto selbstständiger bleibt das Urteil.
Mehr Information ≠ mehr Wissen ≠ mehr Weisheit
Entscheidend bleibt die Fähigkeit, Informationen zu prüfen, einzuordnen und verantwortlich zu nutzen.