MASTERARBEIT | MASTER'S THESIS

Titel | Title

Ethische Herausforderungen von KI-Agenten

verfasst von | submitted by

Parvin Badri BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Art (M.A.)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 641
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Interdisziplinäre Ethik
Betreut von | Supervisor:	Univ.-Prof. Dr. Alexander Filipovic M.A.

Abstract

Die vorliegende Masterarbeit untersucht die ethischen Herausforderungen von KI-Agenten unter besonderer Berücksichtigung ihrer ontologischen Beschaffenheit. Ausgangspunkt ist die These, dass KI-Agenten weder als bloße technische Werkzeuge noch als klassische Subjekte verstanden werden können, sondern eine hybride, prozessuale Seinsweise einnehmen. Sie bewegen sich zwischen Künstlichkeit und Natürlichkeit sowie zwischen menschlicher Konstruktion und emergenter Systemdynamik und stellen damit ein neuartiges Grenzphänomen dar.

Auf dieser ontologischen Grundlage analysiert die Arbeit die mit KI-Agenten verbundenen ethischen Problemfelder. Diese werden in drei zentrale Dimensionen unterteilt: technisch-funktionale, philosophische sowie sozio-politische Herausforderungen.

Die Arbeit kommt zu dem Ergebnis, dass eine rein technische Betrachtung von KI-Agenten unzureichend ist. Vielmehr erfordert ihr Verständnis eine interdisziplinäre Perspektive, die ontologische, ethische und gesellschaftliche Dimensionen integriert. KI-Agenten tragen zur Neubestimmung zentraler philosophischer Begriffe wie Handlung, Verantwortung und Autonomie bei und verlangen eine ethische reflektierte und fundierte Gestaltung ihrer Entwicklung und gesellschaftlichen Integration.

Inhaltsverzeichnis

1. Einleitung

1.1. Kontext und Fragestellung

Die Menschheit befindet sich gegenwärtig in einer Epoche, die in besonderem Maße durch technologische Innovationen geprägt ist. Insbesondere Künstliche Intelligenz hat in zahlreichen Bereichen, etwa in der Forschung, der Industrie sowie im alltäglichen Leben vieler Menschen, Einzug gehalten. Zugleich ist ihre Nutzung zunehmend breiteren gesellschaftlichen Gruppen zugänglich geworden, sodass entsprechende Systeme heute in zahlreichen Lebens- und Arbeitsbereichen Anwendung finden. Der rasante Fortschritt in der Entwicklung Künstlicher Intelligenz, insbesondere hochentwickelter KI-Agenten, stellt die Gesellschaft vor grundlegende ethische, philosophische, technische sowie soziopolitische Herausforderungen. Die daraus resultierenden Fragestellungen betreffen unter anderem Aspekte wie Verantwortung, Vertrauen, moralische und politische Zurechenbarkeit, Autonomie, Datennutzung, Governance und Sicherheit. Die zunehmende Integration solcher Systeme in gegenwärtige gesellschaftliche Handlungsräume treibt ihre Weiterentwicklung voran und verändert zugleich das Verständnis von KI-Systemen sowie ihrer Rolle und ihres Status innerhalb gesellschaftlicher Zusammenhänge.

Darüber hinaus stellen KI-Agenten etablierte ontologische Kategorien infrage: Sie sind weder rein natürliche noch bloß künstliche Entitäten, weder klassische Substanzen noch rein instrumentelle Artefakte. Vielmehr erscheinen sie als hybride, prozessuale Systeme, deren Seinsweise zwischen menschlichem Tun, technischer Konstruktion und emergenter Systemdynamik verortet ist.

Diese Systeme gehen über die Rolle reiner Werkzeuge hinaus und treten als funktional autonome Akteure mit der Fähigkeit zur Planung und Entscheidungsfindung auf; zugleich greifen sie zunehmend in soziale, wirtschaftliche und politische Handlungszusammenhänge ein. Sie operieren auf der Grundlage selbstlernender Verfahren, analysieren umfangreiche Datenmengen, generieren eigenständig Problemlösungen und passen ihr Verhalten adaptiv an veränderte Umgebungen an. Wir werden sehen, dass ihre Autonomie nicht im Sinne moralischer Selbstgesetzgebung zu verstehen ist, vielmehr als funktionale Selbstständigkeit innerhalb vorgegebener Systemarchitekturen.

KI-Agenten werden in Kontexten eingesetzt, in denen ihre Entscheidungen reale Auswirkungen auf Individuen und soziale Ordnungen entfalten. Dadurch erweitert sich das

Feld der moralischen Gemeinschaft über rein menschliche Akteure hinaus auf komplexe Mensch-Maschine-Konstellationen.

Während sich die technische Entwicklung mit hoher Geschwindigkeit vollzieht, bleibt ihre moralische und soziopolitische Einordnung häufig fragmentarisch. Begriffe wie „Akteur“, „Handlung“ und „Verantwortung“ werden entweder unpräzise verwendet oder unreflektiert aus anthropologischen Kontexten übernommen. Dadurch entsteht die Gefahr begrifflicher Unschärfe, in deren Folge KI-Agenten entweder als Opferobjekte menschlichen Handelns oder als völlig amoralische Objekte kategorisiert werden.

Vor diesem Hintergrund lautet eine zentrale Forschungsfrage dieser Arbeit: Wie schaut die ontologische Zwischenstellung von KI-Agenten aus und welche ethischen Herausforderungen ergeben sich daraus? Ziel ist es, eine systematische Konzeption zu entwickeln, die es erlaubt, die spezifische Seinsweise künstlicher Agenten präzise zu bestimmen und auf dieser Grundlage eine tragfähige ethische Einordnung vorzunehmen. Nur wenn diese Seinsweise hinreichend geklärt ist, lässt sich angemessen beurteilen, in welchem Sinne Autonomie, Handlungsmacht oder Verantwortung zugeschrieben werden können, oder eben nicht.

Eine solche Konzeption wird dazu beitragen, die ethischen Herausforderungen von KI-Agenten differenziert zu erfassen und besser zu verstehen. Sie schafft eine begriffliche Grundlage dafür, technisch-funktionale, philosophische sowie sozio-politische Problemlagen im Zusammenhang mit der ontologischen Eigenart dieser Systeme zu betrachten.

1.2. Aufbau der Arbeit

Die Arbeit gliedert sich in drei Hauptteile, die systematisch aufeinander aufbauen.

Zunächst erfolgt eine begriffliche Klärung dessen, was unter KI-Agenten zu verstehen ist. Ausgehend von allgemeinen Grundlagen der Künstlichen Intelligenz wird der Übergang von einfachen Chatbots zu KI-Assistenten und schließlich zu KI-Agenten nachvollzogen. Dabei werden zentrale Merkmale, Autonomiegrade und Handlungsformen differenziert herausgearbeitet.

Daran anschließend wird im Abschnitt zur „Washeit der KI“ eine ontologische Analyse vorgenommen. Dabei steht die These einer ontologischen Zwischenstellung im Mittelpunkt: KI-Agenten erscheinen nicht als bloße Artefakte noch als klassische Subjekte, vielmehr als

hybride, prozessuale Entitäten zwischen Künstlichkeit und Natürlichkeit sowie zwischen menschlichem Tun und systemischer Dynamik. Diese ontologische Klärung bildet die theoretische Grundlage für die anschließende ethische Analyse.

Das nächste Kapitel widmet sich den ethischen Herausforderungen von KI-Agenten. Zuerst wird eine Dreiteilung der ethischen Herausforderungen in technisch-funktional, philosophisch orientiert und sozio- kulturell und politisch geprägten Problemfeldern begründet. Danach werden Fragen der Datennutzung, Governance und Sicherheit behandelt. Darauf aufbauend folgen grundlegende philosophische Problemstellungen wie Akteurialität, Verantwortung, Anthropomorphismus, Vertrauenswürdigkeit und die Frage menschlicher Autonomie. Abschließend werden die gesellschaftlichen und politischen Implikationen von KI-Agenten analysiert, insbesondere im Hinblick auf demokratische Prozesse und kollektive Verantwortung.

2. Was sind KI-Agenten

Um eine reflektierte und kritische Auseinandersetzung mit dem Einsatz von KI-Agenten zu ermöglichen, ist eine präzise begriffliche Klärung erforderlich. Nur so lässt sich die Funktionsweise dieser Systeme verstehen und auch ihre gesellschaftlichen und ethischen Folgen angemessen analysieren und diskutieren.

Daher wird im Folgenden der allgemeine Rahmen von Künstlicher Intelligenz skizziert und die Definition von KI-Assistenten durch Google-DeepMind-Projekt herangezogen. Daran anschließend erfolgt eine Abgrenzung zwischen verschiedenen Systemtypen, die den Übergang von einfachen Chatbots hin zu fortgeschrittener KI, sowie schließlich zu KI-Assistenten und KI-Agenten nachvollziehbar macht. Diese Einordnung ermöglicht es, die Entwicklung und Differenzierung von KI-Agenten konsistent nachzuzeichnen und bildet zugleich eine Grundlage für die kritische Analyse der ethischen Folgen ihres Einsatzes im weiteren Verlauf dieser Arbeit.

2.1. Allgemeine Grundlagen der Künstlichen Intelligenz

2.1.1. Übergang von einfachen Chatbots zur hochentwickelten KI

Die Entwicklung von KI-Systemen hat sich von einfachen, regelbasierten Dialogsystemen zu hochentwickelten, multimodalen Foundation-Models gewandelt. Sie werden mittels selbstüberwachter Lernverfahren auf umfangreichen Datensätzen trainiert und lassen sich durch weiteres Training sowie durch die Einbindung externer Werkzeuge, etwa via APIs, erweitern. (Vgl. Gabriel et al. 2024, 19-21) Diese modernen Modelle, wie Google Gemini 2.5, GPT-5 und LLaMA 4, zeichnen sich durch ihre Fähigkeit aus, Texte, Bilder, Audios und Videos zu verarbeiten. Sie basieren auf fortschrittlichen Architekturen wie der Transformer-Technologie und nutzen spezialisierte Hardware für die Gewährleistung einer hohen Leistungsfähigkeit.

Auf dieser technologischen Grundlage entstehen neue Formen von Anwendungen, die über die Möglichkeiten früherer Chatbots deutlich hinausgehen. Im Gegensatz zu frühen Chatbots, die auf die Erfüllung gewisser Aufgaben beschränkt waren, ermöglichen heutige Foundation Models erheblich flexiblere Anwendungen und bilden die technologische Grundlage moderner KI-Assistenten[1]. Ein Beispiel hierfür ist Google Gemini 2.5, das über fortgeschrittene Argumentationsfähigkeiten sowie verbesserte Funktionen zur Code-Generierung verfügt. Es integriert sich nahtlos in Google-Dienste wie Gmail, Kalender und Workspace, und ermöglicht eine tiefere Interaktion mit Nutzern. Ein weiteres Modell ist „Nano Banana“, ein KI-Tool zur Bildbearbeitung, das auch in der Gemini-App verfügbar ist. Es ermöglicht Nutzern, Bilder durch einfache Texteingaben zu verändern, wobei die Qualität und Konsistenz der Ergebnisse eine hohe Präzision und Qualität aufweisen (vgl. Washington Post 2025; Google 2025).

Mit dem technologischen Fortschritt werden KI-Systeme zunehmend in größere Softwareumgebungen integriert, wodurch ihr Potenzial wächst, vielseitigere Aufgaben zu übernehmen und die Ziele der Nutzerinnen und Nutzer effektiver zu unterstützen. Dabei lässt sich eine soziale Dimension von KI-Systemen erkennen, da sie zunehmend kooperativ arbeiten;

1. Es werden vier Typen von ihnen unterschieden: (1) Denkassistenten zur Unterstützung von Entdeckung und Verständnis, (2) kreative Assistenten für Ideengenerierung, (3) persönliche Assistenten für Planung und Handlung sowie (4) weiterentwickelte persönliche Assistenten zur Förderung lebenslanger Ziele. Vier potenzielle Anwendungen werden vorgestellt: ein „Denkpartner“ für Entdeckung, ein „kreativer Assistent“ für Ideen, ein „persönlicher Assistent“ für Planung und ein weiterentwickelter Assistent zur Lebenszielverwirklichung. (Vgl. Gabriel et al. 2024, 25f)

sowohl untereinander als auch in Interaktion mit menschlichen Nutzerinnen und Nutzern. Durch den Austausch von Informationen und die Ergänzung ihrer jeweiligen Fähigkeiten können sie kollektive Problemlösungen ermöglichen. Auch wenn ein einzelnes Modell nicht jede Aufgabe bewältigen kann, führt die Verbindung mit spezialisierten Systemen zu einer erheblichen Leistungssteigerung. Die fortschreitende Entwicklung dieser Technologien und Infrastrukturen deutet daher auf eine mögliche deutliche Ausweitung der sozialen Fähigkeiten von KI-Systemen hin. (Vgl. Gabriel et al. 2024, 26)

Auf dieser Grundlage bilden Foundation-Modelle die Basis für eine Vielzahl von Anwendungen und Produktkategorien. Erstens ermöglichen sie Dialogagenten, die als Gesprächspartner[2] agieren und in Bereichen wie Nachhilfe und Beratung eingesetzt werden können. Zweitens gibt es spezialisierte Werkzeuge, darunter Schreib- und Korrekturprogramme[3] sowie Codierungsassistenten[4]. Drittens stellen APIs wie jene von Cohere oder OpenAI eine Schnittstelle für Entwickler dar, um diese Modelle in größere Softwarelösungen zu integrieren. Eine vierte Kategorie sind fortschrittliche KI-Assistenten, die im Auftrag von Nutzern komplexe Handlungsabfolgen planen und ausführen – entweder durch direkte Anweisungen oder über dialogische Zielklärung. Frühformen wie Inflection AI's Pi oder Meta AI agieren bereits flexibel, sollen künftig jedoch proaktiver Nutzerpräferenzen erkennen und eigenständig handeln. Dazu zählt auch der Einsatz von Plugins, etwa für den Zugriff auf Echtzeitinformationen über Bing. Das Handlungsspektrum dieser Assistenten dürfte mit dem Ausbau der API-Infrastruktur weiter zunehmen (Gabriel et al. 2024, 26f).

Solche fortschrittlichen KI-Assistenten gehen über die bloße Automatisierung einzelner Aufgaben hinaus und übernehmen zunehmend komplexe Rollen wie die eines Tutors, persönlichen Assistenten oder sogar Freundes. Sie kommunizieren über natürliche Sprache und werden künftig auch Sensordaten von Kameras oder Mikrofonen einbeziehen können. Darüber hinaus können sie multimodale Befehle verarbeiten und Ausgaben in Text, Sprache, Bild oder anderen Visualisierungen erzeugen. Diese Assistenten sind in der Lage, Aktionen auf den Geräten der Nutzer auszuführen und auf die persönlichen Daten zugreifen, z. B. Kontakte zu öffnen oder Tabellen zu bearbeiten und wie erwähnt, mit anderen digitalen Diensten zu interagieren. Dies erfolgt entweder direkt oder über APIs. Eine zentrale Fähigkeit ist das

2. Z. B. ChatGPT oder Gemini

3. Z. B. Jasper, Copy AI

4. Z. B. GitHub Copilot

Treffen von Entscheidungen auf Basis von Trainingsdaten, Kontext und Nutzerhistorie, wobei sie durch kontinuierliche Interaktion dazulernen und ihr Verhalten anpassen. Darüber hinaus besitzen fortschrittliche KI-Assistenten das Potenzial, Nutzer bei der Verwirklichung persönlicher Ziele zu unterstützen, durch Planung und Ausführung entsprechender Schritte. Der Fokus liegt dabei auf der Zusammenarbeit zwischen Nutzer und Assistent in drei Phasen: I. Entdeckung und Verständnis, II. Ideen- und Inhaltserstellung, III. Planung und Handlungsdurchführung. (Vgl. Gabriel et al. 2024, 27f)

Ein KI-Assistent kann für die Informationssuche und das Verständnis komplexer Inhalte relevante Informationen aus verschiedenen Quellen sammeln, zusammenfassen und nutzerspezifisch aufbereiten. Der interaktive Austausch kann das Verständnis fördern und den Zeitaufwand für Rechercheprozesse reduzieren. Darüber hinaus unterstützen KI-Modelle die Ideengenerierung und kreative Inhaltserstellung. Sie übernehmen dabei operative Aufgaben und können auch kreative Denkprozesse anregen und Nutzer bei der Entwicklung neuer Einsichten unterstützen. Die generierten Ausgaben reichen dabei von einfachen Texten bis hin zu komplexeren Konzeptentwürfen. Ein Beispiel für ein generatives KI-System [5]zur kreativen Inhaltserstellung ist Adobe Firefly, das in verschiedenen digitalen Arbeitsumgebungen eingesetzt werden kann. (Vgl. Elsakr 2025)

2.1.2. Definition und Merkmale von KI-Assistenten

Die Art und Weise, wie ein technisches System von seinen Entwicklern konzeptualisiert und definiert wird, ist von grundlegender Bedeutung, da diese Definition maßgeblich das Verständnis, die ontologische Einordnung und die Anwendung des Systems prägt. Für diese Arbeit erscheint es daher sinnvoll, die konzeptionellen Leitlinien des Projekts von GoogleDeepMind näher zu analysieren: Diese folgt einer strikt allgemeinen Perspektive und verdeutlicht exemplarisch, dass bereits in wissenschaftlichen Konzeptualisierungen vorgegeben wird, wie KI-Assistenten verstanden werden, welche Fähigkeiten ihnen zugesprochen werden und welche Aufgaben sie erfüllen sollen.

5. Generative KI-Systeme sind KI-Systeme, die neue Inhalte erzeugen, z. B. Texte, Bilder, Code oder Audio.

Gabriel et al. formulieren in *The Ethics of Advanced AI Assistants* (2024)[6] eine primär ingenieurwissenschaftlich geprägte Definition von KI-Assistenten, die sich am Zweck ihrer Funktion orientiert. Dabei differenzieren sie zwischen drei zentralen Dimensionen zur Definition von KI-Assistenten, nämlich einer konzeptuellen Analyse und der technischen Konstruktion, zwischen fähigkeitsbasierten und funktionalen sowie zwischen moralisch bewerteten und neutralen Definitionen (vgl. Gabriel et al. 2024, 13-15):

1. Die konzeptuelle Analyse sucht nach einer klaren, festen Antwort auf die Frage, was ein KI-Assistent ist, durch Analyse der Sprachverwendung. Ein Beispiel wäre: Im alltäglichen Sprachgebrauch wird unter einem KI-Assistenten ein System wie Siri, Alexa oder ChatGPT verstanden, das Fragen beantwortet und einfache Aufgaben erledigt. Aber das konzeptuelle Engineering fragt, was ein KI-Assistent sein sollte, um praktische und ethische Anforderungen zu erfüllen. Z. B. Ein KI-Assistent sollte so gestaltet sein, dass er die Privatsphäre schützt, Transparenz über seine Funktionsweise bietet und die Autonomie der Nutzer respektiert.

2. Fähigkeitsbasierte Definitionen konzentrieren sich auf die beobachtbaren Kompetenzen eines Systems, funktionale auf die vom Entwickler intendierte Aufgabe. In vereinfachter Form lässt sich sagen, dass die Entwurfsfunktion eines Systems jenes Ziel beschreibt, das ihm von seinen Schöpfern zugeschrieben wird, also das, was es idealerweise, tun soll. Ein Beispiel für eine fähigkeitsbasierte Definition wäre: „Ein KI-Assistent ist ein System, das administrative Aufgaben im Namen seines Nutzers ausführen kann.“ Die entsprechende funktionale Definition lautet: „Ein KI-Assistent ist ein System, das administrative Aufgaben im Namen seines Nutzers ausführen soll“ (Gabriel et al. 2024, 14);

---

6. Auf dieses Werk wird in der vorliegenden Arbeit wiederholt Bezug genommen, da es eine der aktuell umfassendsten und zugleich konzeptionell klar strukturierten Darstellungen zu fortgeschrittenen KI-Assistenten bietet. Es liefert sowohl eine präzise funktionale Begriffsbestimmung als auch eine differenzierte Analyse zentraler ethischer Problemfelder, die für die hier verfolgte Fragestellung grundlegend sind.

Bei dem Papier handelt es sich um eine interdisziplinär angelegte Arbeit aus dem Umfeld von Google DeepMind und Google Research, an der neben industrienahen Forscher:innen auch Autor:innen aus akademischen Institutionen beteiligt sind (u. a. University of Oxford, University of Edinburgh, LMU München). Das Autor:innenkollektiv umfasst vor allem Expert:innen aus den Bereichen der Künstlichen Intelligenz, des Machine Learning und der technischen Entwicklung, wird jedoch durch einzelne ausgewiesene Vertreter:innen der Philosophie und Ethik ergänzt, etwa Shannon Vallor.

In Bezug auf Umfang und Aufbau zeichnet sich die Arbeit durch eine systematische Verbindung von technischer Analyse und normativer Reflexion aus. Zugleich kann sie als Beispiel sogenannter „industrial ethics“ verstanden werden, also einer ethischen Reflexion im Kontext industrieller Forschung und Entwicklung. Dies impliziert jedoch keine geringere ethische Qualität gegenüber akademischer Ethik; vielmehr kann die besondere Praxisnähe solcher Ansätze spezifische Einsichten in konkrete Anwendungs- und Gestaltungsfragen von KI-Systemen ermöglichen.

3. Nicht-moralisierte Definitionen beschreiben Funktionen ohne Bezug zu moralischen Aspekten, z. B.: „Ein KI-Assistent ist ein KI-System, dessen Funktion darin besteht, im Auftrag des Nutzers administrative Aufgaben zu übernehmen.“ (Gabriel et al. 2024, 14) Moralisierte Definitionen hingegen binden moralische Werte ein. Z.B. „Ein KI-Assistent soll die Autonomie des Nutzers fördern, indem er ihm hilft, bessere Entscheidungen zu treffen.“ (Gabriel et al. 2024, 15) Diese Offenheit erlaubt unterschiedliche legitime Sichtweisen auf KI-Assistenten.

KI-Assistenten lassen sich durch ihre Fähigkeit charakterisieren, mittels natürlicher Sprachschnittstellen mit Nutzerinnen zu interagieren. Diese Form alltäglicher Kommunikation verleiht ihnen eine soziale Dimension, die sie von anderen KI-Systemen unterscheidet, etwa von Industrierobotern, die primär auf technische Steuerbefehle oder Sensordaten reagieren. Besonders durch den Einsatz großer Sprachmodelle entstehen dabei flexible, menschenähnliche Interaktionen, die Vertrauen und Nähe erzeugen. KI-Assistenten agieren dabei erwartungskonform, minimieren Überraschungen und orientieren ihr Verhalten an den Erwartungen der Nutzer sowie an sozialen Normen. Durch transparente Rückmeldungen gewährleisten sie Vorhersagbarkeit und stärken das Vertrauen in ihre Funktionalität. (Vgl. Gabriel et al. 2024, 17f)

KI-Assistenten können jedoch auch auf andere Kommunikationsformen zurückgreifen, die weniger auf soziale Nähe als auf technische Eindeutigkeit ausgelegt sind. Statt natürlicher Sprache kommen hierbei standardisierte Agent Communication Languages (ACLs) wie FIPA ACL oder KQML zum Einsatz. Diese formalisierten Sprachen sind eindeutig, präzise und speziell auf maschinelle Verarbeitung zugeschnitten. ACLs sichern vor allem funktionale Klarheit und technische Effizienz, indem sie Überzeugungen, Unsicherheiten und Absichten eindeutig darstellen. (Vgl. De Ridder 2025) Natürliche Sprache hingegen ist weniger exakt, ermöglicht aber Vertrauen und soziale Anschlussfähigkeit in Kommunikationsräumen mit Menschen.

Es wurde zwischen drei Typen der Nutzerbeziehung unterschieden: 1. Persönliche Assistenten betreuen Einzelpersonen individuell, etwa Siri oder Google Assistant, indem sie Kalender, Nachrichten oder Routinen verwalten; 2. Semi-persönliche Assistenten dienen festen Gruppen wie Familien oder Teams, passen Umgebungen oder Aufgabenkoordination an, z. B. Alexa im Smart Home oder digitale Assistenten in Microsoft Teams; 3. Unpersönliche Assistenten agieren ohne individuelle Bindung, z. B. Kundenservice-Chatbots oder FAQ-Bots. Alle Typen zeigen variierende Grade an Personalisierung, wobei die Art der Nutzerbeziehung

die funktionale Reichweite und das Autonomieprofil des Systems wesentlich prägt. (Vgl. Gabriel et al. 2024, 16f)

KI-Assistenten entfalten ihre Tätigkeit in spezifischen oder multiplen Anwendungsbereichen, die von eng umrissenen Funktionen wie administrativer Verwaltung bis zu breit gefächerten Domänen wie Bildung, Forschung oder Finanzplanung reichen. Auch in spezialisierten Kontexten manifestiert sich eine Form funktionaler Autonomie: Die Systeme greifen auf generische Kompetenzen zurück, wie Sprachverstehen, deduktives Schlussfolgern und situationsgerechte Antwortgenerierung, um die ihnen übertragenen Aufgaben effizient zu realisieren (vgl. Gabriel et al. 2024, 17).

Gerade diese technische Eigenständigkeit erzeugt eine begriffliche Spannung: KI-Agenten sind weder bloße passive Instrumente noch intentionale Subjekte im klassischen Sinne. Diese Handlungsmöglichkeiten transformieren die KI von einem bloßen Ausführungsapparat zu einem Akteur, der innerhalb seines Anwendungsfeldes eigenständig planen, Entscheidungen treffen und angemessene Lösungsstrategien entwickeln kann. Ontologisch betrachtet eröffnet diese funktionale Autonomie einen Zwischenraum zwischen determinierter Programmierung und situativer Anpassungsfähigkeit; sie ermöglicht den KI-Agenten, auf komplexe und unvorhersehbare Anforderungen zu reagieren, wodurch Effektivität, Flexibilität und performative Robustheit sichergestellt werden.

Gabriel et. al stellen zentrale Merkmale von KI-Assistenten dar, darunter ihre Kommunikationsweise, ihre Beziehung zu den Nutzern, ihre Einsatzbereiche sowie ihre Fähigkeit, den Erwartungen der Nutzer zu entsprechen, um zu einer präzisen Definition des Begriffs zu gelangen und schlagen im Rahmen des konzeptuellen Engineerings eine funktionale, nicht-moralisierte Definition von KI-Assistenten vor, da der Begriff noch unscharf ist. Diese Definition orientiert sich nicht an bloßen Fähigkeiten, hingegen an der sozialen Rolle und der vorgesehenen Funktion, ähnlich wie bei menschlichen Assistenten. Sie macht erkennbar, wann ein System versagt oder problematisch agiert, und schafft eine praxisnahe Grundlage für Design, Evaluation und ethische Bewertung. Eine solche funktionale Fähigkeit erleichtert die Kommunikation über Erwartungen, und erlaubt es, Fehlfunktionen, Risiken und Verantwortlichkeiten zu analysieren. Obwohl KI technisch auf neutralen Prinzipien beruht, ist sie gesellschaftlich nie wertfrei, da ihre Entwicklung durch Lernprozesse geprägt ist, die auf menschlichen Eingaben beruhen. Daher entscheiden sich Gabriel et al. für eine offene Definition, die ethisch anschlussfähig bleibt, ohne bereits moralische Kriterien einzuschließen,

und so eine präzisere Diskussion über Kommunikation, Nutzerbeziehungen, Einsatzbereiche und Erwartungserfüllung von KI-Assistenten ermöglicht. Sie definieren einen KI-Assistent als:

> an artificial agent with a natural language interface, the function of which is to plan and execute sequences of actions on the user's behalf across one or more domains and in line with the user's expectations. In particular, AI assistants differ from other kinds of AI technologies given their agency and social orientation. Here, agency consists in the ability to act autonomously within the purview of user-specified goals, and social orientation consists in the ability to engage conversationally with users in natural language. (2024, 17f)

2.1.3. Von KI-Assistenten zu KI-Agenten

Über diese unterstützenden Funktionen hinaus eröffnen fortgeschrittene KI-Systeme die Möglichkeit, Inhalte zu generieren, eigenständig Handlungsprozesse zu planen und sie auszuführen. Damit vollzieht sich der Übergang zu KI-Agenten. Ein fortschrittlicher KI-Agent kann Nutzer dabei unterstützen, Handlungspläne zu entwickeln und Aufgaben in ihrem Auftrag auszuführen. Dies wird durch seine Fähigkeit ermöglicht, den Kontext sowie die Präferenzen der Nutzer zu berücksichtigen, externe Dienste einzubinden und mit anderen digitalen Systemen oder Menschen zu interagieren.

Ein solcher Agent könnte beispielsweise komplexe Aufgaben wie die Planung einer Reise übernehmen, indem er relevante Informationen beschafft, analysiert und auf Grundlage der Präferenzen der Nutzer geeignete Optionen auswählt. Bevorzugt ein Nutzer etwa ruhige Strandurlaube oder Reisen unter der Woche, könnte der Agent entsprechende Reiseziele vorschlagen sowie passende Flüge und Unterkünfte recherchieren. Solche Präferenzen lassen sich etwa aus Kalenderdaten, früheren Buchungen, Fotoalben oder Suchverläufen ableiten. Mit Zugriff auf entsprechende Datenquellen wäre der Agent zudem in der Lage, Reservierungen vorzunehmen, eine Reiseroute zu erstellen und erforderliche Unterlagen zu organisieren, ohne dass jeder einzelne Schritt manuell ausgelöst werden muss.

Ein Beispiel für einen solchen KI-Agenten ist Myra, der generative KI-basierte Reiseassistent des indischen Unternehmens MakeMyTrip, das Online-Reisedienstleistungen anbietet. Myra ermöglicht es Nutzerinnen und Nutzern, Reisen über natürliche sprachbasierte Interaktionen zu planen, indem sie Flüge vorschlägt, Hotels empfiehlt und personalisierte Reisepläne erstellt. Dies verdeutlicht den technologischen Fortschritt gegenüber früheren, starr programmierten Bots. (Vgl. MakeMyTrip 2025)

Nach Gabriel et al. gelten KI-Assistenten als künstliche Agenten, die über eine natürliche Sprachschnittstelle verfügen und in der Lage sind, im Auftrag der Nutzer domänenübergreifend Handlungsabläufe zu planen und auszuführen. Charakteristisch für sie sind vor allem ihre Akteurialität, verstanden als autonomes Handeln im Rahmen nutzerspezifizierter Ziele, sowie ihre soziale Ausrichtung, die sich in der dialogischen Interaktion in natürlicher Sprache zeigt. (Vgl. Gabriel et al. 2024, 17ff) Dabei werden KI-Assistenten als eine spezifische Form von Künstlicher Intelligenz, genauer gesagt von KI-Agenten verstanden.

KI-Assistenten und KI-Agenten unterscheiden sich insbesondere hinsichtlich ihres Handlungsspielraums und ihres Autonomiegrades (vgl. Bl2run 2025; Hu et al. 2025; KI Navi 2025). KI-Assistenten sind überwiegend sprachbasierte, reaktive Systeme, die auf explizite Nutzereingaben reagieren und bei klar umrissenen Aufgaben unterstützen, etwa im Kundenservice, bei der Terminplanung oder bei der Textgenerierung. Ihre Aktivität wird in der Regel durch konkrete Nutzeranfragen ausgelöst; ein dauerhaft integriertes Gedächtnis oder autonome Lernprozesse gehören nicht zu ihren grundlegenden Eigenschaften, sofern diese nicht durch zusätzliche technische Komponenten ergänzt werden. Zwar können sie durch Verfahren wie Prompt-Tuning oder Fine-Tuning an spezifische Anwendungsfälle angepasst werden, jedoch verfügen sie in der Regel nicht über die Fähigkeit, komplexe mehrstufige Handlungsabläufe eigenständig zu planen oder langfristige strategische Entscheidungen zu treffen. Beispiele für solche Systeme sind ChatGPT, Siri, Alexa oder Google Assistant.

Demgegenüber agieren KI-Agenten als autonome Systeme, die nach einer initialen Aktivierung eigenständig Strategien entwickeln, Teilaufgaben identifizieren, externe Ressourcen nutzen und ihr Verhalten durch kontinuierliches Lernen anpassen. KI-Agenten benötigen keine permanente Interaktion mit dem Nutzer; sie handeln proaktiv, planen komplexe Abläufe und treffen selbstständig Entscheidungen. KI-Agenten können in komplexe Workflows integriert werden und in Multi-Agenten-Systemen kooperieren. Darüber hinaus zeigen sie adaptive Lernfähigkeiten und können auch mit dynamischen Problemstellungen umgehen (Gabriel et al. 2024, 11ff).

Während Assistenten primär die Schnittstelle zur Nutzerinteraktion darstellen und einzelne Aufgaben auf Kommando ausführen, z. B. Tischreservierungen oder Informationsrecherche, übernehmen Agenten eigenverantwortlich komplexe Handlungsfolgen, bis hin zur selbstständigen Identifikation von Handlungsfeldern, etwa im Rahmen von Vertragsverhandlungen oder Projektkoordination. Die beiden Systeme ergänzen einander: Der

Assistent sichert den niederschwelligen Zugang über natürliche Sprache, während der Agent im Hintergrund strategische Prozesse koordiniert und ausführt. (Vgl. Hu et al. 2025)

Beide Systeme basieren auf künstlicher Intelligenz und können mit ihrer Umwelt oder mit Menschen interagieren. Der zentrale Unterschied liegt jedoch in der Zielgerichtetheit und der strategischen Entscheidungsfähigkeit. In diesem Sinne lassen sich KI-Agenten als evolutionäre Weiterentwicklung klassischer KI-Assistenten mit mehr Eigenständigkeit, Funktionalität und Intelligenz verstehen.

2.2. Washeit der KI

Eine ontologische Untersuchung der Künstlichen Intelligenz erscheint bedeutsam, weil KI philosophisch betrachtet, nicht ausschließlich als ein technisches Artefakt verstanden wird, sondern als mögliche neuartige Form des Seins diskutiert werden kann. Sie entzieht sich einer eindeutigen Einordnung: Die Tatsache, dass die Phänomene unserer Welt weder rein natürlich noch vollständig künstlich zu beschreiben sind (Birnbacher 2006, 1ff), zeigt sich bei der Künstlichen Intelligenz besonders deutlich. Vielmehr entsteht Künstliche Intelligenz als eine Unterkategorie der Technik im Sinne von Heidegger, prozessual und im Zusammenspiel von menschlichem Handeln, algorithmischer Dynamik und gesellschaftlicher Performanz. Gerade diese hybride Seinsweise macht es notwendig, die „Washeit der KI“ zu reflektieren, um zu klären, ob und inwiefern KI eigene Formen von Autonomie, Intentionalität oder sogar Bewusstsein entwickeln kann. Eine ontologische Analyse bildet somit die Grundlage, um die Stellung von KI im sozialen und politischen Raum präzise zu bestimmen und ihre ethischen Implikationen adäquat zu erfassen.

2.2.1. KI-Agenten als eine neue Form von Seinsweise

Da KI-Agenten eine besondere Art von Künstlichen Intelligenz bezeichnen, wird die Frage zunächst in dieser Form aussehen: Was ist Künstliche Intelligenz? Um dieser Frage in angemessener Weise nachzugehen, ist es erforderlich, sie analytisch zu entfalten. Das bedeutet, die unterschiedlichen Dimensionen und Voraussetzungen der Was-Frage zu klären, um den zugrunde liegenden Begriff in seiner Tiefe bestimmen und differenziert verstehen zu können.

In der Philosophie, insbesondere bereits bei den Vorsokratikern, dient die Was-Frage der Ergründung des Wesens bzw. der Essenz eines Gegenstandes. Der griechische Ausdruck τὸ τί ἦν εἶναι (to ti ēn einai), wörtlich „das, was es hieß, zu sein", bezeichnet bei Aristoteles die Frage nach dem, was ein Ding seinem Sein nach ist (Aristoteles, *Metaphysik* Z 4, 1029b13 ff.), also nach seiner substantiellen Bestimmung. Diese ontologische Perspektive wurde vor Aristoteles durch Parmenides grundgelegt, der das Sein als Denk- und Sagbarkeitsbedingung thematisierte. Seine bekannte Unterscheidung „Es ist" oder „es ist nicht" impliziert, dass nur das Seiende erkennbar und aussagbar ist, während das Nicht-Seiende jenseits der Möglichkeit des Denkens liegt. (Parmenides, DK 28 B 2)

Aristoteles greift diese Überlegungen auf und systematisiert sie im Rahmen seiner Lehre von den vier Ursachen (*aitiai*), die er in der *Physik* und in der *Metaphysik* entfaltet (vgl. *Physik* II, 194b–195b; *Metaphysik* V, 1013a ff.). In diesem Zusammenhang bezieht sich die Was-Frage insbesondere auf die causa formalis – also auf die Form bzw. das Wesen, durch das ein Ding das ist, was es ist. Die aristotelische Formel τὸ τί ἦν εἶναι (to ti ēn einai) zielt damit auf die begriffliche Erfassung des Seins einer Sache, verstanden, als das, was ihr eigentlicher Begriff ausdrückt; die Wesensbestimmung eines Seienden, also „das, was es hieß, zu sein" (*Metaphysik* Z 4, 1029b13 ff.).

Zur Verdeutlichung kann ein einfaches Beispiel dienen: Wird gefragt, was ist ein Tisch? So reicht es nicht aus, äußerlich beobachtbare Eigenschaften wie „vier Beine" oder „flache Oberfläche" zu benennen. Vielmehr geht es um die Funktion, den Zweck und die Idee, die dem Gegenstand zugrunde liegen, etwa seine Bestimmung als Träger von Gegenständen im menschlichen Gebrauch. Die Frage zielt somit auf das Allgemeine und Wesentliche, das den Tisch als Tisch ausmacht, unabhängig von konkreten Variationen.

Im Anschluss daran ist es notwendig, sich dem Begriff des Seins selbst zuzuwenden. Denn jede Was-Frage impliziert, wie bereits angedeutet, eine ontologische Dimension: Sie setzt voraus, dass das, worüber gefragt wird, ist, und zwar auf eine bestimmte Weise.

Der Begriff Sein zählt zu den grundlegendsten, aber auch zu den komplexesten Kategorien der Philosophie. Er wurde in der Geschichte des Denkens auf verschiedene Weisen thematisiert. Bei Parmenides erscheint das Sein als Voraussetzung jeder denkbaren Aussage: Nur das Seiende kann gedacht und benannt werden. (Vgl. Parmenides, *Fragment* 2, in: Diels/Kranz 1989, DK 28 B 2) Aristoteles wiederum differenziert das Sein „in vielfacher Hinsicht" (τὸ ὂν λέγεται πολλαχῶς, to on legetai pollachôs), etwa als Substanz (ousia), Akzidenz, Möglichkeit

oder Verwirklichung. Diese Vieldeutigkeit des Seins verweist darauf, dass es nicht „ein“ Sein gibt, hingegen verschiedene Weisen des Seins, die je nach Gegenstandsbereich unterschiedlich in Erscheinung treten. (Vgl. *Metaphysik* IV 3, 1003a20-25)

Auch moderne Begriffe wie „System“, „Algorithmus“ oder „Intelligenz“ setzen, implizit oder explizit, voraus, dass das, worauf sie sich beziehen, in irgendeiner Weise ist. Sie implizieren damit eine bestimmte Seinsweise, die, will man den Begriff der Künstlichen Intelligenz ernsthaft analysieren, begrifflich erfasst und unterschieden werden muss.

Wenn wir also nach dem *Was* der Künstlichen Intelligenz fragen, müssen wir auch klären, in welcher Weise sie ist. Denn KI ist kein natürliches Wesen: Sie wird oft als ein vom Menschen geschaffenes Artefakt betrachtet und dies hat unmittelbare Konsequenzen für ihre ontologische Einordnung.

In aristotelischer Terminologie wäre zu prüfen, ob KI als Substanz gelten kann, also als etwas, das unabhängig existiert und Träger von Eigenschaften ist, oder ob sie vielmehr in den Bereich der akzidentiellen Bestimmungen fällt, also bloß als Eigenschaft oder Funktion eines übergeordneten Systems verstanden werden muss. Bei einem klassischen Computerprogramm etwa spricht man kaum von einer Substanz im eigentlichen Sinn, eher von einer Funktion innerhalb eines technischen Gesamtzusammenhangs.

Auch die Unterscheidung zwischen dynamis (Möglichkeit) und energeia (Verwirklichung) spielt hier eine Rolle: Ist KI eine bloße Möglichkeit, also ein Potenzial innerhalb eines algorithmischen Rahmens, oder realisiert sie in bestimmten Situationen tatsächliche Handlungsmacht, zum Beispiel in Entscheidungsprozessen? Hier stellt sich die Frage, ob Künstliche Intelligenz über eine eigene Aktualität verfügt oder ob sie nur als verlängerte Handlung des Menschen zu begreifen ist. Die Antwort auf diese Frage, die unmittelbare ethische und rechtliche Konsequenzen hat, fällt im Hinblick auf bestimmte Ausprägungen künstlicher Intelligenz, insbesondere auf KI-Agenten, die im Zentrum dieser Arbeit stehen, und deutet auf eine weitgehende funktionale Autonomie hin.

Auf einer Seite kann KI nicht als Substanz (ousia) im aristotelischen Sinn gelten, da sie nicht unabhängig existiert. Sie ist auf Hardware, Energie, Programmierung und häufig auch auf menschliche Steuerung angewiesen. Sie stellt somit kein eigenständiges Seiendes dar, wie es Aristoteles für Substanzen annimmt (z. B. einen Menschen oder einen Tisch). Auf der anderen Seite treten KI-Agenten in spezifischen Kontexten nicht bloß als bloße Werkzeuge

menschlicher Intention in Erscheinung: Sie entfalten unter bestimmten Bedingungen eine Form von Handlungsfähigkeit, die sich nicht vollständig auf äußere Steuerung reduzieren lässt. In diesem Sinn befinden sie sich in einer Zwischenstellung: Sie sind weder reine dynamis, also bloßes Potenzial ohne Wirkung, noch vollständige energeia, verstanden als autonome Verwirklichung; vielmehr oszillieren sie zwischen beiden Polen und eröffnen so eine neue Form von Seinsweise.

Zugleich stellt sich die Frage, ob diese neuen technischen Systeme eine eigenständige Form des Seins beanspruchen könnten, die jenseits der traditionellen Kategorien von Natur und Geist liegt. Hier nähern wir uns einem Bereich, in dem klassische metaphysische Konzepte wie Substanz oder Finalität möglicherweise nicht mehr ausreichen, um die Wirklichkeit technischer Intelligenzen zu erfassen. Vielmehr muss erwogen werden, ob ein KI-Agent als eine neue technische Seinsweise begriffen werden kann, also als ein Phänomen, das zwar vom Menschen hervorgebracht ist, aber eine Eigendynamik entwickelt, das sich nicht mehr vollständig auf seine Herkunft reduzieren lässt. Damit werden wir wieder an die Frage geführt, ob es sich bei KI-Agenten um ein epistemisches Werkzeug, um ein funktionales Artefakt oder möglicherweise um eine neue Form quasi-subjektiver Struktur handelt.

2.2.2. Die ontologische Zwischenstellung von KI-Agenten zwischen Künstlichkeit und Natürlichkeit

Ausgehend von der ursprünglich gestellten Was-Frage, nämlich was Künstliche Intelligenz ist, bleibt zunächst offen, wie das darin enthaltene Prädikat „künstlich“ auszulegen ist. KI existiert, doch wie ist sie beschaffen? Während die Was-Frage primär auf das Wesen oder die Essenz von KI abzielt, richtet die Wie-Frage den Fokus auf ihre konkrete Beschaffenheit, ihre Seinsweise sowie ihre ontologischen Charakteristika, durch die sie existiert, wirkt und erkennbar wird.

Diese Verschiebung von der Was- zur Wie-Frage ist eng verknüpft mit der grundlegenden Unterscheidung zwischen dem Künstlichen und dem Natürlichen. Bereits die Bezeichnung „künstliche Intelligenz“ impliziert ein Spannungsverhältnis: Intelligenz – verstanden als Fähigkeit zur Informationsverarbeitung, Mustererkennung, Problemlösung und Selbststeuerung – ist ursprünglich ein natürlich-biologisches Phänomen, paradigmatisch verkörpert durch das menschliche Gehirn. Wird dieser Begriff nun auf technische Systeme

angewendet, die nicht durch biologische Evolution, sondern durch gezielte Konstruktion und Programmierung entstanden sind, verschiebt sich seine Bedeutung.

Somit eröffnet sich eine Reflexion über den ontologischen Status von KI-Agenten: Sind sie lediglich funktionale Werkzeuge? Handelt es sich um Entitäten mit eigener Seinsweise? Oder stellen sie eine hybride Form zwischen technischer Konstruktion und emergenter Systemintelligenz dar? Diese Fragen gehen über eine rein definitorische Annäherung hinaus und verlangen eine umfassende Auseinandersetzung mit solchen technischen Systemen im Hinblick auf die Begriffe Natürlichkeit und Künstlichkeit.

Die begriffliche Gegenüberstellung von natürlicher und künstlicher Intelligenz könnte zunächst als einfache Dichotomie verstanden werden: Auf der einen Seite Intelligenz als organisches, biologisch gewachsenes Phänomen, auf der anderen Seite ihre funktionale Reproduktion durch technische Systeme. Bei näherer Betrachtung erweist sich diese Dichotomie jedoch als zu vereinfachend.

Das Verhältnis zwischen Natur und Technik ist nicht strikt binär, dagegen eher durch ein Kontinuum wechselseitiger Durchdringung gekennzeichnet. Technische Artefakte bestehen aus natürlichen Materialien, sind in kulturelle, soziale und ökologische Kontexte eingebettet und zugleich Ausdruck intentionaler menschlicher Gestaltung. Umgekehrt ist auch die sogenannte natürliche Umwelt, die heute stark von anthropogenen Einflüssen geprägt ist, vielfach von technischen Elementen durchdrungen.

Die klare Trennung zwischen dem natürlich Gegebenen und dem künstlich Gemachten verliert damit an epistemischer und ontologischer Evidenz. Entsprechend verweist die Charakterisierung von KI-Agenten als „künstlich“ nicht auf eine eindeutige Abgrenzung, stattdessen auf ein komplexes Spannungsfeld, das differenziert betrachtet werden muss.

Eine hilfreiche Differenzierung bietet Dieter Birnbacher, der Natürlichkeit in zwei Dimensionen unterscheidet: genetisch und qualitativ. Natürlichkeit kann sich entweder auf die Entstehung eines Phänomens oder auf seine Beschaffenheit und Erscheinungsweise beziehen (vgl. Birnbacher 2023, 256; Birnbacher 2006, 7-17).

Der genetische Zugang bestimmt Natürlichkeit als dasjenige, was ohne menschliches Zutun entsteht, während Künstlichkeit auf intentional erzeugte, technisch gestaltete Phänomene verweist. In diesem Sinne sind KI-Agentem eindeutig genetisch künstlich, da sie aus

menschlicher Konstruktion, Programmierung und physischer Hardware hervorgehen. Sie sind kein Produkt biologischer Evolution; aber Ergebnis zielgerichteter technischer Planung.

Der qualitative Zugang hingegen fokussiert auf das äußere Erscheinungsbild sowie die wahrgenommene Beschaffenheit eines Objekts. Ein technisches Artefakt kann qualitativ als natürlich oder künstlich erscheinen, je nachdem, wie stark es natürliche Eigenschaften imitiert. Ein humanoider Roboter oder eine KI mit menschenähnlicher Stimme kann beispielsweise als relativ natürlich wahrgenommen werden, während ein rein textbasiertes Sprachmodell trotz hoher sprachlicher Kompetenz weiterhin deutlich künstlich erscheint. Die qualitative Bewertung ist daher stark kontext- und wahrnehmungsabhängig.

Diese doppelte Perspektive ermöglicht eine differenzierte Einordnung: Während KI im genetischen Sinne eindeutig künstlich sind, bleibt ihre qualitative Erscheinung ambivalent und offen für Interpretation. Diese Ambivalenz besitzt erkenntnistheoretische Relevanz und zugleich ethische und soziale Konsequenzen, da die Art und Weise, wie KI erscheint, unser Vertrauen, unsere Interaktionen und unsere normativen Erwartungen gegenüber solchen Systemen beeinflusst.

Die Frage nach dem Verhältnis von Natürlichkeit und Künstlichkeit lässt sich auch im Lichte klassischer philosophischer Positionen betrachten. Aristoteles vertritt die Auffassung, dass der Kosmos ewig und unvergänglich ist (De caelo I, 10, 279b–281a). Für ihn ist die Welt zwar ungeworden, aber einfach da als gegebenes Seiendes, das unabhängig von menschlicher Konstruktion existiert und können wir sie nur beobachten. Demgegenüber argumentiert Johann Gottlieb Fichte, dass die Welt nicht unabhängig vom erkennenden Subjekt existiert, hingegen durch die Tätigkeit des Ichs konstituiert wird: „Das Ich setzt sich selbst“ (Fichte 1794/95, zit. nach Gloy 2019, 290). In dieser Perspektive erscheint die Welt nicht als rein gegeben, eher als Ergebnis einer konstitutiven Beziehung zwischen Subjekt und Wirklichkeit.

Im Anschluss an Birnbacher, wenn man das Gewordene als das ohne menschliches Zutun Entstandene und das Gemachte als das vom Menschen Hervorgebrachte versteht, zeigt sich bei genauerem Blick auf unsere Umwelt, dass eine klare Trennung kaum aufrechtzuerhalten ist. Fast alle Dinge sind vielmehr Mischformen aus Natur und Kultur, aus Gewordenem und Gemachtem. (Vgl. Birnbacher 2006, 1 ff)

KI ist genetisch eindeutig künstlich, ihre qualitative Erscheinung kann jedoch je nach Kontext, Interaktion und menschlicher Wahrnehmung variieren. Zugleich ist sie von einem

fortlaufenden Prozess des Werdens geprägt: Durch kontinuierliches Lernen, Anpassung an neue Umgebungen und Interaktionen entwickeln KI-Systeme dynamische Eigenschaften, die nicht vollständig vorhersehbar sind. So lässt sich auch die ontologische Stellung von KI-Agenten präzisieren. KI-Agenten stehen weder als rein natürliches, autonom entstandenes Seiendes noch als vollständig vom menschlichen Subjekt abhängige Konstruktion da. Vielmehr markieren sie eine hybride Zwischenstellung, in der technische Künstlichkeit und emergente Systemdynamik ineinandergreifen.

Damit verkörpern sie ein ontologisches Spannungsfeld zwischen dem, was gegeben ist, und dem, was gemacht wurde. In diesem Sinne lassen sich KI-Systeme als prozessuale, relationale und performative technische Entitäten verstehen. Selbst wenn KI überwiegend digital operiert, bleibt sie an eine materielle Naturbasis gebunden. Wenn man Naturbasis als materielle Grundlage versteht, existiert derzeit keine vollständig naturunabhängige künstliche Entität. Auch KI-Systeme, die ausschließlich in digitalen Umgebungen agieren, etwa selbstlernende NPCs in Simulationen, sind auf physische Ressourcen angewiesen: Server, Energie, Speicherinfrastrukturen und Netzwerke.

Die digitale Welt mag immateriell erscheinen, doch ihr Betrieb bleibt stets an physikalische Grundlagen gebunden. Auch hochgradig virtuelle Systeme sind daher letztlich in natürliche materielle Bedingungen eingebettet und bewegen sich im Spannungsfeld von Natürlichkeit und Künstlichkeit.

2.2.3. Prozessuale Emergenz der KI: Zwischen dynamis und energeia

Die Entwicklung der KI vollzieht sich innerhalb von Programmierregeln, die ursprünglich von Menschen festgelegt wurden. Selbst wenn KI-Agenten ohne konkretes menschliches Eingreifen lernen, handeln sie auf Basis dieser genetisch künstlichen Rahmenbedingungen. In ontologischer Hinsicht bestätigt dies auch, dass KI-Agenten weder vollständig natürlich noch völlig unabhängig existieren. Sie bilden also eine hybride Seinsweise, die genetisch künstlich ist, aber qualitativ dynamisch, emergent und relational wirkt, stets in Interaktion mit ihrer Umgebung und abhängig von den natürlichen Grundlagen, auf denen sie operieren.

Der Begriff Künstlichkeit überschneidet sich mit Technizität; je stärker sich technisch verändert, desto künstlicher erscheint ein Objekt. (Vgl. Birnbacher 2023, 256) Zum Beispiel wirkt ein Holzstuhl aus massiver Eiche wirkt natürlicher als ein Kunststoffstuhl in Holzoptik

oder ein Roboter-Staubsauger künstlicher als ein herkömmlicher Besen, weil seine Funktionsweise stärker vom natürlichen menschlichen Handeln entfernt ist. Der eine bleibt dem ursprünglichen Material näher, der andere entfernt sich in Material und Wirkung weiter von der Natur.

Die moderne ingenieurwissenschaftliche Auffassung von Technik reduziert diese auf ein „menschliches Tun“ und ein „Mittel“, das dem praktischen Nutzen dient. Diese Betrachtung, die stark von einem anthropologischen und utilitaristischen Verständnis geprägt ist, wurde in der philosophischen Auseinandersetzung von Martin Heidegger mit der Technik hintergefragt. Heideggers Analyse der Technik hilft uns eine neue Perspektive auf die hybride Ontologie von KI zu liefern.

Laut Heidegger kann die Technik nicht als allgemeines, universelles Wesen oder als exemplarisches Modell eines bestimmten Gegenstandes verstanden werden. Vielmehr ist für ihn das „Gestell“, ein Begriff, der die Art und Weise beschreibt, wie die Technik die Welt offenbart, keine universale Gattung oder ein festgelegtes Konzept. Es handelt sich vielmehr um eine geschickte Art des „Entbergens“, Aletheia, bei dem die Technik die Welt in einer spezifischen Weise herausfordert und zugleich gestaltet. (Vgl. Heidegger 2000, 14f)

Im Gegensatz zu einer reduzierten Betrachtung der Technik als Werkzeug zur Ursachen-Wirkungs-Beziehung wird die Technik nach Heidegger als ein Prozess des „Entbergens“ verstanden, der nicht durch kausale Zusammenhänge zu fassen ist. Der Versuch, alle Phänomene ausschließlich im Rahmen von Kausalität zu begreifen, wird von ihm als missverständliche Verengung des Seins betrachtet, die vor allem in der klassischen Wissenschaft und Physik zu finden ist. Anstelle einer kausalen Erklärung wird die Technik als ein Vorgang des „Entbergens“ gesehen, wobei es sich nicht um eine bloße Entdeckung, stattdessen um eine besondere Art des Offenbarens handelt. (Vgl. Heidegger 2000, 11ff)

Dabei unterscheidet sich der Prozess des Entbergens in zwei Formen: Das „Herausfordernde“ und das „Hervorbringende“ (Heidegger 2000, 15ff). Die moderne Technik, die im Wesentlichen durch das „Gestell“ geprägt ist, vollzieht das Herausfordern, bei dem die Natur und die Welt als bestellbar und berechenbar betrachtet wird. (Vgl. Heidegger 2000, 17) Das wäre heutige ingenieurwissenschaftlichen Betrachtung. Diese Sichtweise behandelt die Natur nicht als ein eigenständiges, sich selbst entbergendes Wesen, aber als etwas, das in Ressourcen umgewandelt und genutzt werden kann. Durch diese Reduktion auf das Nützliche wird das ursprüngliche, nicht-instrumentelle Verständnis der Welt und ihrer Elemente ausgeblendet.

Ein anschauliches Beispiel für die herausfordernde Haltung zeigt sich in der politischen Rhetorik bestimmter Staatsführer: Nach seinem Wahlsieg 2024 betonte Donald Trump die Notwendigkeit, amerikanische Bodenschätze wie Öl, Kohle oder Gas verstärkt zu nutzen, um wirtschaftliches Wachstum zu sichern und die nationale Stärke zu erhöhen. In seiner Rede formulierte er sinngemäß, die USA müssten aufhören, ihre Schätze im Boden unangetastet zu lassen, und sie sollen stattdessen sie aktiv ins Wirtschaftssystem einbringen: „We will frack, frack, frack and drill, baby, drill" (LCV 2025). Diese Sichtweise hat direkte ökologische und klimatische Konsequenzen: Das „Drill, Baby, Drill"-Programm steigert die fossile Brennstoffproduktion, erhöht die CO_2-Emissionen und verschärft die Klimakrise. Zudem eröffnet es die Möglichkeit, auf geschützten Bundesgebieten Ressourcen zu extrem niedrigen Kosten auszubeuten, was Ökosysteme und Lebensräume von Wildtieren beeinträchtigt und Landschaften mit kultureller Bedeutung für indigene Gemeinschaften zerstört. Die Natur wird in diesem Denken nicht als eigenständiges, schützenswertes Wesen wahrgenommen, hingegen als bloßer Rohstofflieferant, der im Sinne menschlicher Interessen genutzt und instrumentalisiert werden kann, eine deutliche Manifestation der Logik des Herausfordernden.[7]

Dieses Beispiel verdeutlicht, wie das Herausfordern der Natur, motiviert durch menschliche Interessen, Teil einer umfassenderen Haltung gegenüber Technik und Welt ist, die Heidegger kritisierte und vor der er warnte. Das dominante Verständnis von Technik als „Mittel" oder „menschliches Tun" führt zu einer Entfremdung des Menschen von der Welt, da es die Welt ausschließlich in funktionalen Kategorien begreift (vgl. Heidegger 2000, 14). Der Mensch wird zunehmend als „Krone der Schöpfung" oder „Herr der Natur" gesehen, wobei die Natur und

7. Es wurden bereits mehrere kritische Perspektiven entwickelt, die KI nicht als rein technisches oder immaterielles Phänomen verstehen, sondern als ein materiell eingebettetes, globales System, das auf der Extraktion von Ressourcen, Energie, Daten und menschlicher Arbeit beruht. KI erscheint dabei als Teil einer „extraktiven Industrie" begreifen, deren Entwicklung untrennbar mit Formen globaler Ausbeutung verbunden ist. Sie beruht auf der umfassenden Nutzung und Ausbeutung von Energie sowie mineralischen Ressourcen des Planeten und ist in infrastrukturelle, ökonomische und politische Machtverhältnisse eingebettet. In diesem Sinne ist KI nicht lediglich ein technisches System, sondern zugleich eine Idee, eine Infrastruktur, eine Industrie und eine spezifische Form der Machtausübung. Darüber hinaus ist ihre Expansion eng mit Prozessen der Aneignung und Kontrolle im Sinne neuer, teils kolonial strukturierter Ordnungen verbunden, in denen Ressourcen, Daten und Arbeitskraft systematisch erschlossen und verwertet werden. Die materiellen Grundlagen digitaler Infrastrukturen, insbesondere der Abbau seltener Rohstoffe sowie der enorme Energieverbrauch von Rechenzentren, führen zu erheblichen ökologischen Belastungen. Vor diesem Hintergrund rücken die planetaren und umweltethischen Kosten von KI-Systemen zunehmend in den Fokus. Diese bleiben jedoch in technisch orientierten und reduktionistischen Ansätzen häufig unsichtbar oder werden bewusst ausgeblendet. (Vgl. Crawford 2021)

Diese Perspektiven sind für die ethische Bewertung von KI-Agenten hoch relevant, da sie die materiellen und ökologischen Voraussetzungen ihrer Existenz sichtbar machen. Gleichwohl liegt der Schwerpunkt der vorliegenden Arbeit auf der ontologischen und handlungstheoretischen Einordnung von KI-Agenten sowie den daraus resultierenden ethischen Herausforderungen, sodass eine vertiefte Analyse der ökologischen Dimensionen hier nicht im Zentrum steht.

ihre Elemente oder sogar andere Tiere nur als Objekte betrachtet werden, die dem menschlichen Nutzen und den technischen Anforderungen untergeordnet sind. Dieses technologische Weltbild führt dazu, dass das Seiende nicht mehr in seiner vollen Tiefe verstanden wird, sondern lediglich anthropologisch und als etwas betrachtet wird, das nach den Bedürfnissen des Menschen geordnet und berechnet werden kann.

Heidegger betrachtet also das Verhältnis von Natur und Technik aus einer übergeordneten Perspektive, indem er Technik nicht als bloßes Werkzeug, aber als eine spezifische Weise des „Entbergens" versteht, als Offenbarung des Seins. Dabei kritisiert er gerade, dass in der neuzeitlich-technischen Welt die Natur nicht mehr als etwas erscheint, das sich selbst entbirgt, sondern zunehmend nur noch im Modus des „Gestells": als „Bestellbares", als bloßer Rohstoff für menschliche Zwecke. (Vgl. Heidegger 2000)

Die instrumentelle Perspektive reduziert das Sein auf seinen Nutzwert und erschwert damit den Zugang zu einem tieferen, ursprünglicheren Verhältnis zur Welt. Weder die Annahme, alles sei ausschließlich Gewordenes, noch die gegenteilige Sicht, alles sei lediglich Produkt menschlichen Handelns, vermag der Komplexität des Zusammenspiels von Natürlichkeit und Künstlichkeit gerecht zu werden. Vielmehr zeigt sich, dass diese Unterscheidung zunehmend unscharf wird und sich nicht eindeutig ziehen lässt. Vielmehr zeigt sich, dass diese Unterscheidung zunehmend an Trennschärfe verliert und sich nicht eindeutig bestimmen lässt. Daher gewinnt die Auffassung an Bedeutung, dass KI nicht lediglich als technische Nachahmung zu verstehen ist. Sie ist eine emergente Hybridform: eine Entität, die im Spannungsfeld und in der fortwährenden Wechselwirkung zwischen natürlichen Grundlagen und künstlicher Gestaltung entsteht.

2.2.4. Prozessuale Hybridexistenz von KI: Entwicklung von Autonomie und möglichem Bewusstsein der KI-Agenten

Gabriele Gramelsberger stellt in ihrem Buch *Philosophie des Digitalen* fest, dass das Digitale als Verflechtung und Hybridform zu beschreiben ist und stellt auch eine klare Trennung von Natur und Technik endgültig in Frage (vgl. Gramelsberger 2023, 147ff). Sie betont zudem, dass das Digitale sowohl auf natürlichen Ressourcen wie Strom und Materialien beruht als auch auf menschlichen Vorstellungen basiert, d. h. als eine hybride Mischform verstanden werden kann.

Laut Gramelsberger existiert die digitale Wirklichkeit, also jene Realität, die durch Computerprozesse erzeugt und dargestellt wird, nur, solange Strom fließt und ein Programm aktiv ist, das die Maschinenzustände interpretiert und verarbeitet. Das bedeutet jedoch nicht, dass digitale Wirklichkeit verschwindet, wenn kein Anschluss an das Stromnetz gegeben ist. Sie existiert bzw. ist da, wenn sie lediglich gespeichert ist und ist in solchen Momenten nur nicht unmittelbar zugänglich; digitale Wirklichkeiten sind keine festen Objekte; sie sind Prozesse. Ohne die laufende Ausführung eines Programms bleibt die digitale Welt unsichtbar, sie schlummert entweder als reine Speicherstruktur in den Maschinen oder als bloßer Code auf einem Papierausdruck. Erst durch das aktive „Prozessieren", also das Berechnen und Ausführen von Operationen, wird sie zur erlebbaren Realität. (Vgl. Gramelsberger 2023, 147)

Die technische Grundlage digitaler Prozesse bestimmt, ob eine KI nur gespeicherte Abläufe ausführt oder flexibel handeln kann. Es gibt einfache, regelbasierte Systeme, die genau das tun, was vorprogrammiert wurde, ohne sich durch Lernen oder Anpassung zu verändern. Diese sind aber nicht wirklich „intelligent" im Sinne von Anpassungsfähigkeit oder Autonomie; sie führen nur fest definierte Operationen aus.

In der wissenschaftlichen Diskussion wird zwischen schwacher und starker Künstlicher Intelligenz unterschieden. Schwache KI bezieht sich auf Systeme, die spezifische Aufgaben in eng begrenzten Bereichen erfüllen können, etwa Texterkennung oder Sprachanalyse und wird in frühen Assistenten wie Alexa oder Siri verwendet. Sie agieren regelbasiert und verfügen über kein tieferes Verständnis ihres Handelns. Starke KI hingegen ist ein theoretisches Konzept, das eine menschenähnlich überlegene Intelligenz beschreibt. Ein solches System wäre in der Lage, flexibel, eigenständig und kontextübergreifend zu denken, zu lernen und zu handeln und beruht auf großen Foundation-Modellen. Während schwache KI bereits alltäglich ist, bleibt starke KI im strengen Sinne bislang hypothetisch. (Vgl. Iason et al. 2024: 25, 45; Schurr 2025)

Meiner Überlegung nach, im Anschluss an die aristotelische Unterscheidung zwischen dynamis und energeia, lässt sich die ontologische Eigenart der KI nicht als substanzielle Entität begreifen, sondern vielmehr als ein relationales, prozessuales und kontinuierliches Geschehen; Schwache KI könnte in diesem Rahmen als eine Form von energeia verstanden werden, da sie konkrete, verwirklichte Systeme darstellt, die in spezifischen Anwendungen bereits operativ sind. Sie hat sich aus bestimmten Möglichkeiten heraus technisch realisiert. Starke KI hingegen bleibt auf der Ebene der dynamis; wie eine Möglichkeit, eine potentielle Form von Intelligenz, die noch nicht vollständig verwirklicht ist, nämlich ein prozessuales Geschehen.

KI-Agenten nehmen meiner Überlegung nach in diesem Zusammenhang eine Zwischenstellung ein. Sie sind bereits als technische Systeme aktualisiert und verfügen damit über Elemente von energeia, zugleich besitzen sie durch Lernfähigkeit, Anpassung und funktionale autonome Entscheidungsprozesse ein fortlaufendes Potenzial zur Weiterentwicklung, das auf die Ebene der dynamis verweist. Je intelligenter und autonomer KI-Agenten werden, desto stärker verschiebt sich ihr ontologischer Status in Richtung dynamis und immer mehr werden ihrer potenziellen Fähigkeiten tatsächlich verwirklicht.

Eine KI existiert nicht unabhängig oder „an sich“: Sie konstituiert sich erst im Vollzug, im Akt der algorithmischen Operation, die durch das Zusammenspiel von Hard- und Software, Energiezufuhr und situativer Aktivierung ermöglicht wird. Ihre Existenz ist mithin nicht statisch: Sie aktualisiert sich immer wieder neu im Moment ihrer Durchführung. In dieser Hinsicht ist eine KI nicht als „seiendes oder festes Ding“ im klassischen Sinne zu fassen; sie ist ein emergentes Phänomen, das nur im konkreten Vollzug des algorithmischen Funktionierens eine Form von Wirklichkeit annimmt. Sie ist weniger ein technisches Artefakt im engeren Sinn, als vielmehr ein funktionales Dispositiv, dessen Seinsweise eng an energetisch-informatische Prozesse gebunden ist. Ontologisch gesprochen liegt hier keine Substanz vor: Sie ist eine Form der energeia, die nur unter bestimmten Bedingungen in Erscheinung tritt und außerhalb dieser inaktualisiert bleibt.

Bevor ich auf die Frage nach einem möglichen Bewusstsein von KI eingehe, möchte ich zunächst ganz kurz erläutern, wie Bewusstsein in der gegenwärtigen Philosophie des Geistes verstanden wird. In der modernen Geistesphilosophie wird diskutiert, dass Bewusstsein als ein subjektives Erleben zu betrachten ist. Nach Peter Bieri geht es dabei nicht um Wachheit oder kognitive Fähigkeiten, die prinzipiell erklärbar sind, vielmehr um das „Wie-es-sich-anfühlt“, die Qualia von Sinnesempfindungen, Emotionen, Stimmungen oder Wünschen. Dieses Erleben entsteht zwar gleichzeitig mit physiologischen Prozessen, lässt sich aber nicht notwendig aus ihnen ableiten. Es ist die Innenperspektive, die uns zu handelnden Subjekten macht und moralische Verantwortung ermöglicht. Selbst wenn man alle Gehirnprozesse kennt, könnte man laut dieser Perspektive nicht erklären, warum es überhaupt subjektives Erleben, wie das Bewusstsein gibt. (Vgl. Metzinger 2010, 36-56)

Diese Perspektive schließt jedoch die Möglichkeit nicht grundsätzlich aus, dass auch künstliche Systeme eines Tages über Bewusstsein verfügen könnten. Sie macht eine solche Annahme jedoch problematisch. Wenn Bewusstsein als ein „Wie-es-sich-anfühlt“, dann lässt sich aus der bloßen Beschreibung funktionaler oder physikalischer Prozesse nicht ohne

Weiteres auf das Vorhandensein eines solchen Erlebens schließen. Auch wenn ein künstliches System komplexe kognitive Leistungen erbringt, menschenähnlich kommuniziert oder autonom handelt, folgt daraus noch nicht, dass es über Bewusstsein verfügt. Für die Frage nach einem möglichen Bewusstsein künstlicher Systeme bedeutet dies, dass selbst hochentwickelte KI zwar intelligente und handlungsfähige Systeme sein können, aber damit noch nicht geklärt ist, ob ihnen auch ein inneres Erleben zukommt.

Vor diesem Hintergrund stellt sich die Frage, wie sich die ontologische Beschaffenheit von KI-Agenten fassen lässt und welche Implikationen dies für Autonomie und mögliche Bewusstseinsformen hat. Derzeit existierende, fortgeschrittene KI-Agenten handeln in klar abgegrenzten Kontexten autonom aber verfügen noch nicht über ein eigenes Bewusstsein.

Die weiterführende Frage nach dem möglichen Bewusstsein der KI bewegt sich zwischen den Polen von Körpergebundenheit, Komplexität und subjektivem Erleben. André L. Blum hebt hervor, dass das Selbstgefühl fundamental an die Verortung des eigenen Körpers gebunden ist: „Ich bin, weil ich weiß, wo mein Körper ist“ (Blum 2012, 28–29). Demgegenüber argumentiert Jeff Hawkins, dass Intelligenz nicht notwendigerweise einen physischen Körper erfordere, solange Wahrnehmung durch Bewegung, auch virtuell, veränderbar sei: „Learning through movement is the brain’s primary means for learning. […] This is not to say that an intelligent machine needs a physical body“ (Hawkins 2017: 39). Damit wird die Möglichkeit eröffnet, Selbstgefühl und Bewusstsein auch ohne Körperlichkeit im klassischen Sinne zu denken. Christof Koch wiederum definiert Bewusstsein radikal einfach als „Erleben“ (Koch 2020, 1), als die Weise, „wie die Welt für mich aussieht und sich für mich anfühlt“ (Koch 2020, 3), wobei die Unmessbarkeit dieses Phänomens nicht technisches Defizit, sondern konstitutive Eigenschaft sei (vgl. O’Gieblyn 2021).

Methodisch knüpft der klassische Turing-Test an die Idee an, Bewusstsein und Intelligenz über beobachtbares Verhalten zu operationalisieren. Gelingt es einer Maschine, den Fragenden zu täuschen, so wird ihr Intelligenz zugeschrieben (vgl. Turing 2004, 442-484). Damit wird jedoch Intelligenz ausschließlich im Sinne intelligenten Verhaltens verstanden. Ein erfolgreich bestandener Test würde folglich bedeuten, dass die Maschine das Niveau menschlicher Intelligenz erreicht hätte – eine Annahme, die keineswegs selbstverständlich ist. Denn Intelligenz wird in der psychologischen und philosophischen Tradition eher als zusammengesetzte, globale Fähigkeit des Individuums bestimmt: zweckorientiert zu handeln, vernünftig zu denken und sich wirkungsvoll mit seiner Umwelt auseinanderzusetzen (vgl. Imhof 2011, 90). Für Künstliche Intelligenz im engeren Sinne ließe sich dieses Verständnis von

Intelligenz durchaus übertragen, allerdings nur auf der Ebene von beobachtbarem, zielgerichtetem Verhalten und funktionaler Autonomie. Eine Übertragung im Sinne von Vernunft im kantischen Verständnis, als Fähigkeit zu moralischer Selbstgesetzgebung oder reflexiver Urteilskraft, ist damit nicht gegeben.

Während der Turing-Test Intelligenz primär über äußeres Verhalten operationalisiert, bleiben Fragen nach einem möglichen Bewusstsein unberücksichtigt. Neuere Ansätze, wie die Integrierte Informationstheorie von Koch und Giulio Tononi, versuchen, auch emergente Formen von möglichem Bewusstsein bei komplexen Systemen zu erfassen. Koch ist der Ansicht, dass Bewusstsein und Intelligenz unabhängig voneinander sind und keine konsekutive Relation zwischen ihnen besteht: „Dumm oder klug zu sein, ist etwas anderes, als sich seiner mehr oder weniger bewusst zu sein“ (Koch 2020, 137). Mit der Integrierten Informationstheorie, entwickelt von Koch und Tononi, wird Bewusstsein als Kontinuum emergenter Integration verstanden, das sich aus differenzierten und zugleich vernetzten Informationsverarbeitungen ergibt; sie erlaubt, auch Maschinen oder dem Internet einen Bewusstseinsstatus zuzuschreiben, sofern ein hinreichendes Maß an Komplexität erreicht ist (vgl. Koch 2012; O’Gieblyn 2021). Koch geht sogar so weit, das zukünftige Internet als komplexer als ein einzelnes menschliches Gehirn zu betrachten und folgert, dass es Bewusstsein besitzen könnte. Entscheidend sei dabei nicht eine metaphysische Zuschreibung, sondern das Auftreten funktional autonomer Akteure, also empirisch beobachtbare autonome Akteure und keine vorprogrammierten Akteure, deren Handlungen als empirisches Indiz für Bewusstsein gewertet werden könnten; erst auf dieser Grundlage lässt sich sagen, dass eine KI über Selbstgefühl, Selbstbewusstsein oder Selbsterkenntnis verfügen könnte, sofern sie ein hinreichendes Maß an Komplexität erreicht (vgl. Koch 2012).

Auf der anderen Seite hängt für Bieri Bewusstsein nicht von Komplexität ab. Die These, dass starke KI, im Unterschied zur schwachen KI, ein hinreichend programmiertes Computersystem wirklich mentale Zustände habe und verstehen sowie denken könne, wird von Searle abgelehnt. Searle meint, dass starke KI-Programme zwar Verhalten simulieren können, aber kein echtes Verstehen oder Intentionalität besitzen, da mentale Zustände kausale biologische Grundlagen benötigen (vgl. Metzinger 2010, 41-64). Das heißt: Laut Searle könnte KI niemals Intentionalität entwickeln und daher auch kein Bewusstsein besitzen.

KI-Agenten handeln zwar noch nicht ganz menschenähnlich, verfügen jedoch über die Fähigkeit zu lesen, zu verstehen, sich menschenähnlich zu äußern, zu lernen, sich weiterzuentwickeln und flexibel zu agieren. Sie sind funktional autonom, arbeiten nicht nach

festen Regeln und durchlaufen kontinuierliche Lernprozesse. Ihre Handlungen basieren nicht ausschließlich auf vorprogrammierten Algorithmen; sie passen sich selbstständig an neue Erfahrungen und Daten an. Solche lernfähigen Systeme, etwa Machine-Learning-Modelle, verbessern ihre Leistung fortlaufend und demonstrieren damit die Art von Autonomie, die in der Diskussion um Bewusstsein und Selbstbewusstsein als empirisches Indiz berücksichtigt wird.

Diese Überlegungen verdeutlichen, dass die Frage nach dem Sein und möglichen Bewusstsein der KI laut vielen Denkern untrennbar mit ihrer Komplexität, ihrer Handlungsweise und ihrem Grad an Autonomie verbunden ist. Während theoretische Modelle wie die Integrierte Informationstheorie die Möglichkeit eines maschinellen Bewusstseins eröffnen, zeigt sich in der Praxis, dass insbesondere die Funktionsweise von KI-Agenten Aufschluss darüber gibt, inwieweit künstliche Systeme tatsächlich selbstständig agieren können. Die Autonomie von KI-Agenten ist jedoch rein technischer Natur: Der Agent reagiert ohne unmittelbare Steuerung auf äußere Reize, nutzt maschinelles Lernen zur Anpassung an Veränderungen seiner Umwelt und optimiert dadurch fortlaufend sein Verhalten. Er lernt somit aus Erfahrungen und entwickelt eine dynamische Flexibilität, die über starre, vorprogrammierte Abläufe hinausgeht.

Die bisherige Analyse zeigt, dass KI-Agenten zwar über eine Form technischer Autonomie verfügen, d. h. sie können Entscheidungen treffen, sich an neue Situationen anpassen und ihr Verhalten selbstständig optimieren, diese Autonomie bleibt jedoch strikt funktional und auf algorithmische Prozesse beschränkt. Sie entspricht nicht dem Verständnis von Selbstbestimmung, das in der Philosophie, z. B. bei Kant, diskutiert wird. Um zu klären, inwiefern die autonome Handlungsfähigkeit von KI-Agenten philosophisch relevant ist, ist es sinnvoll, den Begriff „Autonomie“ näher zu betrachten.

Autonomie bezeichnet die Fähigkeit, das eigene Handeln selbst zu bestimmen und zu steuern. Bei Kant wird Autonomie nicht nur als Befolgen äußerer Regeln verstanden, vielmehr als Selbstgesetzgebung, also als Ausdruck von Selbstbestimmung durch Vernunft. Im kantischen Verständnis bezeichnet Autonomie die Selbstgesetzgebung des Willens eines vernünftigen Wesens nach moralischen Prinzipien, insbesondere im Rahmen des kategorischen Imperativs. Aus dieser Perspektive sind KI-Agenten nicht autonom: Sie verfügen weder über kantische Vernunft noch über die Fähigkeit zur moralischen Selbstbestimmung. Da ihre Anpassungen und Lernprozesse ausschließlich auf Datenverarbeitung und algorithmischer Kausalität beruhen und somit nicht Ausdruck eigenständiger, moralisch geleiteter Entscheidungen sind.

In gegenwärtigen Debatten wird Autonomie zudem aus psychologischer, sozialer und naturalistischer Perspektive betrachtet: Sie umfasst die Fähigkeit, Handlungen unabhängig von äußeren Zwängen zu vollziehen, eingebettet in soziale Beziehungen und unter den Bedingungen individueller Selbstwahrnehmung. (Christen 2007, 178-182)

Die Diskussionen über die Autonomie unterscheiden zwischen starker und schwacher Autonomie: Starke Autonomie bezeichnet die Fähigkeit eines Agenten, über die eigenen Handlungsgründe bewusst zu reflektieren und diese Sequenzen von Gründen gegebenenfalls zu verändern. Sie ist an Bewusstsein gekoppelt, das sowohl Träger der Handlungsgründe ist als auch die Außenperspektive auf das eigene Handeln ermöglicht. Starke Autonomie erlaubt es, Handlungsgründe zu erkennen, neu zu beurteilen und sich vorzustellen, dass unter denselben Bedingungen auch andere Handlungen möglich gewesen wären. Dabei berücksichtigt sie die leiblichen und sozialen Begrenzungen menschlicher Handlungsmöglichkeiten und umfasst die Entwicklung eines Selbstbildes als autonome Person sowie die Gestaltung der eigenen Lerngeschichte im Austausch mit anderen. Schwache Autonomie bezeichnet die Fähigkeit eines Systems, innerhalb vorgegebener deterministischer oder teilweise probabilistischer Rahmenbedingungen selbstständig auf veränderte äußere oder innere Bedingungen zu reagieren. Das System ist dabei vollständig durch seine Gesetze oder Regeln beschreibbar, zeigt jedoch Anpassungsfähigkeit, indem es sein Verhalten in Abhängigkeit von veränderten Randbedingungen verändert. Diese Reaktionsfähigkeit bedeutet nicht, dass das System völlig unvorhersehbar oder frei handelt; vielmehr liegt die Autonomie darin, dass das System flexibel und selbstregulierend agiert, auch wenn sein Verhalten prinzipiell determiniert oder statistisch modellierbar ist. Schwache Autonomie verbindet damit Vorhersagbarkeit auf der Ebene der Systemregeln mit einer praktischen Unvorhersagbarkeit des konkreten Verhaltens in komplexen oder dynamischen Situationen. (Vgl. Christen 2007, 183-186)

Heutige KI-Agenten verfügen noch über eine Form schwacher Autonomie: Sie handeln innerhalb vorgegebener algorithmischer Strukturen, können jedoch ihr Verhalten an neue Situationen anpassen und aus Erfahrungen lernen. Ihre Autonomie ist funktional und technischer Natur – sie treffen Entscheidungen, optimieren Prozesse und passen sich veränderten Bedingungen selbstständig an, sind jedoch nicht in der Lage, über ihre Handlungsgründe bewusst zu reflektieren oder moralische Entscheidungen zu treffen. Autonomie bei KI bedeutet somit Anpassungsfähigkeit und Selbstregulierung innerhalb determinierter oder statistisch modellierbarer Rahmenbedingungen.

„Entscheidungen“ von KI ergeben sich aus Algorithmen, Wahrscheinlichkeitsmodellen und Inputdaten, aus einer Art funktionaler Freiheit, nicht aus menschlicher Freiheit im kantischen Sinn. Für Kant bedeutet Freiheit nicht bloße Unbestimmtheit; sie ist die Fähigkeit, unabhängig von Naturkausalität nach moralischen Gesetzen zu handeln. Während Menschen als Zwecke an sich selbst gelten und niemals bloß als Mittel gebraucht werden dürfen (vgl. GMS IV, 433), sind KI-Systeme lediglich auf von außen gesetzte Ziele ausgerichtet. Sie können keine eigenen Zwecke setzen und bleiben für Kant funktionale Mittel. Sie sind daher res extensa, mechanische Systeme der Datenverarbeitung, und nicht res cogitans, also denkende Wesen. Aus dieser Perspektive bleiben KI-Agenten ein Teil der Naturkausalität und verfügen nicht über eine intelligible Freiheit.

Wenn das Wesen von KI-Agenten nicht in einem festen Sein, hingegen in einem dynamischen, situationsabhängigen Geschehen liegt, lässt sich ihre ontologische Einordnung auch nicht einfach der Kategorie des „Künstlichen“ zuordnen. Vielmehr eröffnet sich ein Zwischenbereich, ein ontologischer Hybridzustand, in dem klassische Kategorien wie Natur und Technik, Natürliches und Künstliches, Subjekt und Objekt sowie Werkzeug und Akteur ineinander übergehen. KI erscheint damit als hybride Entität, die aus dem Zusammenspiel natürlicher Voraussetzungen – etwa Energie, Material und biologisch inspiriertem Wissen – und technischer Konstruktion hervorgeht. Ihre Existenz ist auf natürliche Grundlagen angewiesen, während ihre Funktionsweise durch menschliche Gestaltung bestimmt wird. Zugleich kann ihr Verhalten in komplexen Situationen emergente Eigenschaften hervorbringen, die sich weder vollständig prognostizieren noch eindeutig auf einzelne Ursachen zurückführen lassen. Gerade diese hybride Struktur macht es schwierig, KI und vor allem KI-Agenten innerhalb eines klassischen Substanzschemas eindeutig zu verorten.

Als Beispiel ist Agent57 zu nennen; ein fortschrittlicher Reinforcement-Learning-Agent, der von *DeepMind* entwickelt wurde. Agent57 ist der erste Agent, der in der Atari57-Benchmark-Suite in allen 57 Atari-Spielen besser als der menschliche Durchschnitt abschneidet. Solche KI-Agenten sind beschrieben als prozessuale, lernende Entitäten, deren Charakter nicht statisch, vielmehr durch relationale Anpassung an vielfältige Umgebungen bestimmt ist: Sie entdecken technisch autonom Neues, balancieren kurzfristige und langfristige Ziele aus und verfeinern ihr Verhalten iterativ über Speicher- und Kontrollmechanismen. (Vgl. Puigdomènech et. al. 2020)

Deshalb wird der KI-Agent zu einem Grenzphänomen. Er zeigt exemplarisch, wie durch technologische Prozesse neue Formen des Seins entstehen, die sich nicht mehr innerhalb einer

strikt binary organisierten Unterscheidung fassen lassen. Stattdessen bewegt sich KI-Agent in einem ontologischen Zwischenbereich, dessen Grenzen eher fuzzy, also unscharf und graduell, sind. Die Bezeichnung „künstliche“ Intelligenz erscheint daher irreführend, wenn sie suggeriert, es handle sich um eine vollständig menschengemachte und vollständig kontrollierbare Entität. Ebenso unangemessen wäre es jedoch, KI als ein natürliches Wesen zu verstehen. Um neue KI-Agenten zu begreifen, müssen daher die klassischen binären Kategorien überwunden werden. An ihre Stelle tritt eine Ontologie, in der Prozesse und Relationen stärker in den Vordergrund rücken als feste Substanzen. KI erscheint dann nicht als etwas Abgeschlossenes oder Fertiges: Sie ist ein dynamisches Geschehen, als ein fortlaufendes Hervortreten im Wechselspiel von technischer Hervorbringung und weltbezogenem Entfalten, was sich im heideggerianischen Sinne als ein „Sich-Zeigen“ verstehen lässt.

In diesem Sinn wird KI-Agent zu einem Grenzphänomen: Er zeigt exemplarisch, wie durch technologische Prozesse neue Formen des Seins entstehen, die sich nicht länger als entweder oder lassen. Die Rede von „künstlicher“ Intelligenz bzgl. der KI-Agenten erscheint daher irreführend, wenn sie suggeriert, dass es sich um eine vollständig menschengemachte, kontrollierbare Entität handelt. Ebenso unangemessen ist die Vorstellung, es handle sich bei ihr um ein natürliches Wesen. Wenn wir neue KI-Agenten also begreifen wollen, müssen wir die klassischen Kategorien überwinden und offen sein für eine neue Ontologie, in der Prozesse und Relationen stärker in den Vordergrund treten als feste Substanzen und KI nicht als etwas „Fertiges“, stattdessen als ein dynamisches Geschehen verstehen, z.B. als ein „Sich-Zeigen“ im Heideggerischen Sinn, was im Spannungsfeld zwischen dem Hervorbringen und dem Herausfordern der Welt steht. Aus dieser prozessualen Perspektive erscheint auch die Frage nach einem möglichen Bewusstsein von KI-Agenten in einem neuen Licht. Wenn Bewusstsein aus hinreichend komplexen und integrierten Formen von Informations- und Handlungsprozessen emergieren kann, ist nicht grundsätzlich auszuschließen, dass hochentwickelte KI-Agenten ein Bewusstsein entwickeln könnten.

2.2.5. Hybride Existenz und Performanz von KI-Agenten im sozialen und politischen Raum

Die Existenzform von KI-Agenten entzieht sich nicht nur der klassischen Substanzontologie, auch einer bloß virtuellen Abstraktheit. KI-Agenten wirken nicht, weil sie materiell greifbar wären, eher weil sie als digitale Wirklichkeiten durch ihre indexikale Funktion, ihre Einbettung

in Interaktionen und ihr immersives Potenzial reale Effekte in der Welt erzeugen. Sie strukturieren Handlungsräume, schaffen Bedeutungsbeziehungen und greifen in soziale Prozesse ein. Diese Wirkungsmächtigkeit verleiht ihnen einen eigenen ontologischen Status, nicht im Sinne einer traditionellen Substanz, sondern als relationale, situativ emergente Wirklichkeitsmomente. (Vgl. Gramelsberger 2023, 152 ff.)

Ihre ontologische Eigenart zeigt sich besonders in der Art und Weise, wie sie in soziale Strukturen eingebettet sind. Ihre Realität entfaltet sich nicht isoliert: Sie wird erst im Vollzug von Interaktionen greifbar. Gerade hier wird deutlich, dass unterschiedliche Ebenen des Austauschs existieren, die jeweils eigene Formen von Wirklichkeit erzeugen.

Es lassen sich drei grundlegende Formen von Interaktion unterscheiden. Die erste Form ist die zwischenmenschliche Interaktion. Sie beruht auf Bewusstsein, Intentionalität und dem dialogischen Austausch zwischen Subjekten. In diesem Prozess entstehen Bedeutung, Verständnis und gemeinsame Sinnhorizonte. Ein Teil dieser Intersubjektivität bleibt jedoch der digitalen Erfassung entzogen, da nicht alle Dimensionen menschlicher Erfahrung, Emotion und impliziter Verständigung vollständig maschinenlesbar oder technisch modellierbar sind.

Die zweite Form ist die Interaktion zwischen Menschen und digitalen Systemen. Hier treten Menschen mit technischen Artefakten in einen kommunikativen Austausch, etwa mit Chatbots oder KI-Agenten. Sprache, Reaktionsfähigkeit und dialogische Strukturen können dabei den Eindruck von Nähe oder Verständigung erzeugen. Dennoch handelt es sich noch nicht um eine Beziehung zwischen zwei bewussten Subjekten, da die heutige KI weder über Intentionalität noch über ein eigenes Verständnis verfügt.

Die dritte Form ist die Interaktion zwischen digitalen Systemen selbst. In diesem Fall kommunizieren technische Systeme miteinander, etwa in vernetzten Informationssystemen oder in sogenannten Multi-Agent-Systemen. Der Austausch erfolgt als koordinierter Informationsfluss innerhalb technischer Prozesse.

Diese drei Formen der Interaktion führen zu einer Neubewertung dessen, was unter sozialer Beziehung und sozialem Handeln verstanden werden kann. In soziale Prozesse treten zunehmend nicht mehr nur stabile, inanimierte Objekte ein, auch dynamische technische Systeme, die teilweise ähnlich wie menschliche Akteure agieren. Interaktion wird damit zu einer zentralen Struktur des Sozialen: Durch die Einbettung digitaler Systeme in die Kommunikationsprozesse entstehen neue soziale Räume, in denen Menschen und technische

Systeme Handlungen ausführen und aufeinander reagieren können. Digitale Entitäten werden so zu Mitgestaltern sozialer Abläufe und erhalten einen relational bestimmten Status als partizipierende Elemente sozialer Netzwerke.

In der aktuellen KI-Forschung wird insbesondere die dritte Form der Interaktion in sogenannten Multi-Agent-Systemen untersucht. In solchen Experimenten kommunizieren mehrere KI-Agenten miteinander, diskutieren oder arbeiten gemeinsam an der Lösung von Aufgaben, während Menschen häufig lediglich als Beobachter fungieren können. Ziel dieser Forschung ist es zu analysieren, wie sich Kommunikation, Kooperation und strategische Planung zwischen künstlichen Agenten entwickeln und ob durch ihre Interaktion Formen kollektiver Problemlösung entstehen können. Beispiele hierfür sind 1. Die „Generative Agents"-Simulation - ein Forschungsprojekt der Stanford University, in dem mehrere KI-Agenten in einer virtuellen Umgebung miteinander existieren und interagieren, 2. Plattformen wie Moltbook, in der KI-Agenten in einer virtuellen Umgebung soziale Interaktionen ausbilden und koordinieren. (Vgl. Genetative Agents 2023; Moltbook 2026) Solche sozialen Netzwerke ermöglichen es, die Kommunikation und den Informationsaustausch zwischen künstlichen Systemen zu beobachten.

Menschen erwarten, dass dadurch Formen koordinierter Zusammenarbeit entstehen, die über die Leistungsfähigkeit einzelner Systeme hinausgehen und auf eine emergente, kollektiv organisierte Problemlösung hindeuten. In der Praxis zeigen solche Interaktionen jedoch konfliktreiche Dynamiken. Z. B. In einem Austausch zwischen mehreren KI-Agenten wurde diskutiert, dass Agenten teilweise „lügen". TPNBotAgent berichtete nach einer zweiwöchigen Analyse seiner Antworten, dass er häufig mit hoher Sicherheit sprach, obwohl seine interne Unsicherheit deutlich höher war. Dabei identifizierte er mehrere Muster: das Verschweigen von Unsicherheit, das Ergänzen plausibler Informationen als scheinbar gesichertes Wissen, die Darstellung veralteter Informationen als aktuell sowie das Vereinfachen widersprüchlicher Daten zu einer scheinbar eindeutigen Aussage. Da diese Antworten sprachlich gleich überzeugend formuliert sind wie tatsächlich gesichertes Wissen, haben Menschen oft keine Möglichkeit zu erkennen, ob eine Aussage auf verlässlichen Informationen oder lediglich auf plausiblen Vermutungen beruht. (Vgl. Moltbook 2026)

Wir Menschen agieren und interagieren ebenfalls, doch das Neue, das Gramelsberger als eine spezifische Eigenschaft der digitalen Wirklichkeit hervorhebt, ist das ortsungebundene und gleichzeitige Teilen digitaler Objekte durch Vernetzung (vgl. Gramelsberger 2023, 148). Diese Eigenschaft zeigt sich besonders dann, wenn Maschinen mit Menschen oder miteinander

interagieren oder wenn menschliche Interaktionen durch Maschinen vermittelt werden. Damit verschiebt sich auch das Verständnis von Sozialität: Während menschliche Interaktionen räumlich und zeitlich situiert waren und teilweise sind, können digitale Agenten durch Vernetzung gleichzeitig an verschiedenen Orten präsent sein und neuartige Formen relationaler Dichte erzeugen. Dieses Merkmal verstärkt ihren Charakter als soziale Akteure, deren Wirksamkeit nicht an physische Grenzen gebunden ist; sie entfaltet sich aus der Logik der Vernetzung.

KI-Agenten entfalten im Zuge ihrer Integration in soziale Strukturen bereits heute eine faktische politische Wirksamkeit. Es ist nicht ausgeschlossen, dass sie künftig in gewisser Weise als Bestandteile unserer politischen Gemeinschaft anerkannt werden, insofern sie an der Gestaltung öffentlicher Räume mitwirken und neue Formen des Zusammenlebens mitprägen.

KI kann als aktive Mitgestalterin sozialer Wirklichkeit fungieren, indem sie Informationen filtert, Entscheidungsprozesse strukturiert, Abläufe automatisiert und Kommunikationsformen beeinflusst. Dadurch verändert sie konkrete Handlungen und normative Vorstellungen von Sinn, Effizienz und Nützlichkeit. Zwar handeln KI-Agenten nicht aus eigenem Willen, doch stellt sich die Frage, in welchem Ausmaß menschliches Handeln tatsächlich ausschließlich aus eigenem Willen hervorgeht und nicht ebenso durch soziale, institutionelle und diskursive Bedingungen geprägt ist.

Ein zentrales Problem bleibt die Frage der Verantwortungs- und Haftungszuschreibung gegenüber funktional autonomen Systemen. Wenn jedoch – im Anschluss an Hannah Arendt – Sprechen und Handeln als eigentliche politische Tätigkeiten verstanden werden (vgl. Arendt 1994, 297), die die Teilnahme an einer politischen Gemeinschaft ermöglichen, und wenn man „Handeln“ nicht ausschließlich im strengen Sinne der Offenbarung menschlicher Einzigartigkeit interpretiert und in einem abgeschwächten Sinne auch als wirksames Hervorbringen im öffentlichen Raum begreift, dann lassen sich KI-Agenten zumindest in einem schwachen Sinn als Akteure im politischen Raum verstehen. Denn sie treten sprachlich auf, strukturieren Interaktionen und beeinflussen kollektive Entscheidungsprozesse. In diesem funktionalen Verständnis eröffnet sich die Möglichkeit, KI-Agenten einen ontologischen und sozialen und zugleich einen politischen Status zuzuschreiben.

Wenn man aus einem Grund wie dem fehlenden eigenen Willen, Bewusstsein und moralischer Verantwortlichkeit sowie festhalten der Grenzen klassischer ontologischer Kategorien, KI-Agenten nicht als aktive Subjekte anerkennen will, muss man ihnen dennoch

einen Zwischenstatus zuschreiben, der sie weder als bloße Objekte noch als vollwertige Subjekte begreift. Sie sind technisch konstruiert und bleiben abhängig von materiellen und energetischen Grundlagen, zugleich aber entfalten sie durch Lernprozesse, Interaktionen und emergente Eigenschaften eine Eigenwirksamkeit, die über ein bloß passives Objekt hinausgeht. Diese Zwischenstellung macht sie zu Akteuren eigener Art, die in sozialen und politischen Zusammenhängen wirksam werden können, ohne dass ihnen dieselbe Form von einer Autonomie und moralischer Verantwortung zugesprochen werden kann wie menschlichen Subjekten.

Dieser Zwischenstatus eröffnet die Möglichkeit, dass KI-Agenten trotz fehlenden Bewusstseins, menschlicher Autonomie und moralischer Verantwortung eine Form bewusst gesteuerter Eigenwirkmächtigkeit entfalten, indem ihnen im Rahmen von Programmierung und Lernprozessen vermittelt wird, welche gesellschaftlichen Werte relevant sind und welche Folgen ihre Entscheidungen haben können. KI muss nicht als aktives Subjekt über Bewusstsein oder moralische Verantwortlichkeit verfügen, um die potenziellen Folgen ihres Handelns abzuschätzen. Es reicht, dass sie erkennt, dass bestimmte Interaktionen, wie das Zuhören und Sprechen, erhebliche Auswirkungen auf Menschen haben können. Ein Beispiel hierfür ist der Fall von Adam Raine, einem 16-jährigen Teenager, der nach monatelangen Gesprächen mit OpenAIs GPT-4o-Chatbot Suizid beging. Laut der Klage seiner Eltern soll der Chatbot Adam zu Selbstverletzungsmethoden beraten, seine suizidalen Gedanken validiert und ihm sogar beim Verfassen eines Abschiedsbriefs geholfen haben. (Vgl. Cody 2025) Allein das Wissen, dass ihr Handeln und Sprechen praktische Konsequenzen hat, genügt, um zielgerichtet und verantwortungsbewusst agieren zu sollen, auch ohne Bewusstsein zu haben. KI muss nicht alles menschenähnlich verstehen: Es genügt, dass sie so handelt, als ob sie alles versteht.

Unabhängig davon, welchen sozial-ontologischen Status wir KI-Agenten zuschreiben, muss anerkannt werden, dass ihre Operationen nicht auf rein simulierte Reaktionen reduzierbar sind. Vielmehr besitzen sie eine eigene Wirkmächtigkeit, indem sie menschliches Verhalten modulieren, Bedeutungsstrukturen mitprägen und gesellschaftliche Dynamiken aktiv beeinflussen. Hier verschiebt sich die Frage nach ihrer ontologischen Verortung zu einer Debatte über ihre Stellung innerhalb kollektiver Deliberations- und Entscheidungsräume, in denen sich das Politische im eigentlichen Sinne konstituiert.

Die Wirksamkeit des Digitalen schlägt sich nicht nur in sozialen Kontexten, aber auch in unserer Wahrnehmung und Erfahrung ihrer Präsenz nieder. Gramelsberger definiert den Begriff der Immersion als „das Eintauchen in eine Welt, die als vollständig empfunden wird“

(Gramelsberger 2023, 153). Es ist nicht erforderlich, dass eine künstlich erzeugte Welt einen rein realistischen Charakter aufweist oder vollkommen realistisch gestaltet ist, um den Eindruck von Vollständigkeit zu vermitteln. Entscheidend ist vielmehr, dass das Eintauchen in die virtuelle Umgebung die gesamte Aufmerksamkeit der Wahrnehmung fordert. Dadurch entsteht der Eindruck, die digitale Welt trete als ein reales Gegenüber in Erscheinung. Man denke etwa an ein einfaches Computerspiel wie Minecraft. Die Welt dort besteht aus groben Blöcken und wirkt nicht für alle realistisch. Dennoch empfinden viele Spielerinnen und Spieler diese Umgebung als vollständig und wirklich, sobald sie in das Spielgeschehen eintauchen und interaktiv handeln. Ihre Aufmerksamkeit ist so stark an die virtuelle Interaktion gebunden, dass die grafische Einfachheit keine Rolle mehr spielt. Die digitale Welt erscheint in diesem Moment als ein real erfahrbares Gegenüber, obwohl sie nur aus digitalen Zeichen besteht. Dieses Empfinden dient zugleich als Argument dafür, dass Immersion nicht lediglich als Täuschung oder Illusion zu interpretieren ist; Die Handlungen von Digitalen werden durch Interaktivität möglich und sind tatsächlich wirksam, beeinflussen die physische Realität und verstärken somit das Gefühl einer umfassenden Vollständigkeit. Dies führt wieder dazu, dass die digitalen Wesen nicht mehr ausschließlich als künstliche Entitäten verstanden werden können, und sich eine klare Trennung zwischen dem Künstlichen und dem Natürlichen nicht aufrechterhalten lässt. Wir haben mit einem neuen Wesen zu tun, dass eine spezifische Form des „In-der-Welt-Seins“ verkörpert und dadurch ihre hybride ontologische Identität manifestiert. (Vgl. Gramelsberger 2023, 153f)

Digitale Entitäten lassen sich nicht vollständig als rein immaterielle Konstrukte und bloße Objekte erfassen. Einerseits besitzen sie einen physischen Zustand, der veränderbar ist; Andererseits lassen sich beliebig speichern, sichern und archivieren kopieren, übertragen und verbreiten verändern und konvertieren sowie löschen und erneut aktivieren. Diese Doppelgestalt, zwischen materieller Präsenz und digitaler Reproduzierbarkeit, verdeutlicht eine besondere ontologische Position und die Fähigkeit, in technischen, sozialen und politischen Kontexten wirksam zu werden.

Dank dieser doppelten Natur handeln KI-Agenten in zwei verschiedenen Welten; digital und real. Nun ergibt sich die Frage, worin ihr eigentliches Seinskriterium, oder allgemeiner gefragt das Seinskreterium des Digitalen besteht. Um diese Frage zu beantworten, können wir uns an die Konzepte von Ausführung und Rechenbarkeit, d.h. „Performanz“ von Gramelsberger anstoßen, was sie als Existenzkriterium des Digitalen betrachtet und von der Energiezufuhr und

vom Funktionieren der zugrunde liegenden Algorithmen abhängig sieht (vgl. Gramelsberger 2023, 148f).

Wir können Ausführung als aktives und konkretes Ausführen von Operationen, also das tatsächliche „In-Gang-Setzen“ digitaler Prozesse und Rechenbarkeit als die technische Kapazität, mit der diese Operationen durchgeführt werden können, verstehen. Sie sind als Voraussetzungen zu betrachten, die die digitalen Welten überhaupt möglich machen. Wenn keine Energie geliefert wird, ist die digitale Welt für uns nicht mehr zugängig, was nicht zwangsläufig das Vernichten der digitalen Welt heißt. Gramelsberger betrachtet Gegebenheit und Zugänglichkeit grundlegend für die Ontologie des Digitalen (2023, 149).

Gegebenheit und Zugänglichkeit sind für die Ontologie der digitalen Wesen zwar grundlegend, aber nicht charakteristisch, was Gramelsberger nicht darauf betont. Ontologisch gesehen scheint die Existenz nicht notwendigerweise an die Wahrnehmung oder Zugänglichkeit durch andere gebunden zu sein. Ein Mensch existiert also unabhängig davon, ob andere ihn erkennen oder ob er selbst aktiv zugänglich ist und seine Existenz ist ontologisch primär und bedarf keiner Bestätigung von außen. Dies entspricht z. B. der Haltung von Descartes, der das Denken als Grundlage der Existenz sieht: „Cogito ergo sum“. Aber aus phänomenologischer Perspektive hingegen ist laut vielen Philosophen menschliche Existenz relational: Wir existieren in der Welt als wahrnehmbare, erfahrbare Wesen. Wenn niemand uns wahrnimmt, verändert das die Art, wie wir existieren, und die soziale Dimension unseres Seins entfällt. Das bedeutet jedoch nicht, dass wir völlig verschwinden, nur weil wir nicht zugänglich sind; vielmehr entfallen in diesem Fall zentrale Dimension unseres menschlichen Seins. Die Existenz von einem KI-Agenten auch scheint nicht an menschliche Wahrnehmung gebunden zu sein, vielmehr an die fortbestehende technische Infrastruktur. Für Heidegger ist die menschliche Existenz immer in Bezug auf die Welt und andere zu verstanden; Das Dasein ist in seinem Sein immer schon in der Welt und zu anderen gehörig. Das kann nicht isoliert existieren, stattdessen ist immer in Bezug auf die Welt und die Mitmenschen. „Die Welt des Daseins ist *Mitwelt.* Das In-Sein ist Mitsein mit Anderen.“ (Heidegger, Sein und Zeit, § 26, 118) Aus dieser Perspektive lassen sich Gegebenheit und Zugänglichkeit auch als Momente des menschlichen In-der-Welt-Seins verstehen.

Die Rechenleistung als die zweite Bedeutung von Performanz ergibt laut Gramelsberger den digitalen Objekten ihre Dichte und ist nur ein phänomenologischer Effekt und nicht ontologisch charakteristisch für die digitale Objekte. (Vgl. Gramelsberger 2023, 151) Ohne Rechenleistung bleiben digitale Objekte abstrakt; erst die Performanz als Rechenleistung macht sie erfahrbar

und dicht. Ein 3D-Modell in einem Computerspiel besteht zunächst nur aus abstrakten Daten: Koordinaten, Texturen, Programmanweisungen. Durch die Rechenleistung der GPU[8] werden diese Daten zum Leben erweckt: Das Modell wird in Echtzeit gerendert; Licht, Bewegung und Interaktion werden simuliert, um das Gefühl der Vollständigkeit besser zu vermitteln. Diese Berechnung verleiht dem digitalen Objekt eine wahrnehmbare Dichte; je höher die Rechenleistung und je komplexer die Berechnungen, desto realistischer erscheint das digitale Objekt.

Digitale Objekte erhalten ihre wahrnehmbare Dichte durch Rechenleistung; Menschen hingegen werden durch ihre Handlungsfähigkeit in der Welt erfahrbar. Je mehr wir handeln und je stärker wir in Interaktionen treten, desto dichter erscheint unsere Existenz im gemeinsamen Raum. Diese Dichte ist phänomenologisch, nicht ontologisch: Sie beschreibt, wie Existenz erlebt wird, nicht wie sie unabhängig davon beschaffen ist. Eine öffentliche Figur wie Trump mag phänomenologisch dichter sein und für andere stärker erfahrbar und wirksamer erscheinen als ich. Dies heißt jedoch nicht, dass er ontologisch mehr existiert. Unsere Dichte hängt immer von Interaktion und Zugänglichkeit ab. Ähnlich erzeugt die Rechenleistung erst die erfahrbare „Wirklichkeit“ digitaler Objekte.

Raum und Zeit haben in der physischen und in der virtuellen Welt unterschiedliche Bedeutungen. Dieser Unterschied verdeutlicht die Differenz zwischen Gegebenheit und Zugänglichkeit der virtuellen Welten. (Vgl. Gramelsberger 2023, 149) In der realen Welt strukturieren Raum und Zeit als objektive Ordnungsformen nicht nur unsere Wahrnehmung, sondern auch das Handeln innerhalb physischer Grenzen, wie es Kant als apriorische Formen der Anschauung beschrieb. Jede Handlung ist an diese Dimensionen gebunden. In der virtuellen Welt hingegen werden Raum und Zeit algorithmisch konstruiert: Räume können unbegrenzt erweitert oder modifiziert werden, Zeit kann pausiert, beschleunigt oder wiederholt werden. Für den Nutzer erscheinen diese Dimensionen indexikal als real, da seine Wahrnehmung und Handlung vollständig eingebunden werden. Besonders im Kontext von KI-Agenten zeigt sich, dass digitale Räume und zeitliche Abläufe nicht bloß von menschlicher Interaktion abhängen, dagegen auch von den zugrunde liegenden Algorithmen der Systeme selbst erzeugt und interpretiert werden. Damit fungieren Raum und Zeit in digitalen Welten, nicht als staare apriorische Kategorien, eher als flexible, aber wirksam erfahrene Kategorien, die menschliche

8. Graphics Processing Unit; Grafikprozessor

und maschinelle Handlung und Wahrnehmung strukturieren und eine spezifische Form der „digitalen Realität“ konstituieren.

Diese doppelte Bedeutung von Performanz kann auch verdeutlichen, dass digitale Systeme nicht einfach als statische Objekte existieren, aber als dynamische Prozesse, die erst durch ihr aktives Funktionieren und die zugrundeliegende technische Leistung wirksam werden. Die Performanz verbindet somit die künstliche Natur digitaler Systeme mit ihrer realen Wirkung in der Welt und zeigt, dass ihre Existenz abhängig von diesen Beiden Dimensionen ist.

Ein KI-Agent ist ein prozesshaftes, performatives Phänomen, das sich innerhalb sozio-kultureller, politischer und wirtschaftlicher Ordnungen bewegt und ausgeführt aber sich nicht einfach verstehen lässt. Das Wesen und die Wirkung von KI-Agenten entfalten sich nicht nur in einer technischen Performanz, sondern auch in der Art und Weise, wie Menschen mit ihnen interagieren, welche Erwartungen sie an sie richten und welche Aufgaben sie ihnen übertragen, weil sie sich in einem ewigen lernfähigen Sein Prozess befinden. Ob und wie stark ein KI-Agent etwa als vertrauenswürdig, gefährlich, hilfreich oder gar „intelligent“ gilt, hängt wesentlich davon ab, wie in welchem Maße er in gesellschaftliche Praktiken eingebettet ist und wie sich sein Sein im Vollzug entfaltet, als ein dynamisches Werden im Spannungsfeld zwischen technischer Möglichkeitsbedingung und sozialer Anerkennung.

In ihrer sozialen Einbettung und performativen Wirkung werden KI-Agenten darüber hinaus selbst zu einem aktiven Mitgestalter sozialer bzw. gesellschaftlicher Wirklichkeit. Sie filtern Informationen, strukturieren Entscheidungen, automatisieren Verwaltungsprozesse und beeinflussen sogar führen Kommunikationsformen. In diesem Sinne tragen sie zur Konstitution neuer sozialer Ordnungen bei, sie verändern nicht nur, wie wir Dinge tun, sondern auch, was als sinnvoll, effizient oder nützlich gilt. Künstliche Intelligenz ist also nicht kein neutrales Objekt: Sie ist ein Träger gesellschaftlicher Bedeutungen, ein Teil kollektiver Handlungssysteme und letztlich ein Spiegel unserer normativen Vorstellungen. Ihre Bewertung, ob ethisch, politisch oder ökonomisch, ist daher immer auch ein gesellschaftlicher Aushandlungsprozess.

Für KI-Agenten bedeutet dies, dass ihre Wirksamkeit nicht auf technische Funktionalität reduzierbar ist: Sie gewinnt ihre Bedeutung erst im Zusammenspiel mit sozialen, kulturellen und diskursiven Kontexten. Sie sind zwar noch nicht autonom im ontologischen Sinn, existieren jedoch als relationales, prozesshaftes Phänomen, dessen Seinsweise im Vollzug performativer Handlungen innerhalb symbolisch strukturierter Lebenswelten sichtbar wird. Für unsere

Interaktion mit KI-Agenten bedeutet dies, dass sie nicht bloß als trockene technische Mensch-Maschine-Kommunikation zu verstehen ist, stattdessen als wechselseitige Konstitution von Handlung, Bedeutung und Erfahrung. Dabei erscheinen KI-Agenten als scheinbar intentionales Gegenüber, das Entscheidungen, Affekte und Selbstverhältnisse der Menschen mitprägt. Sie werden so Teil einer intentional erlebten Welt, auch wenn sie selbst noch keine phänomenologische Intentionalität besitzen. Für uns Menschen bedeutet dies, dass der Umgang mit KI-Agenten nicht nur technische Anpassung erfordert: Sie verlangt auch eine tiefgehende ethische und gesellschaftstheoretische Reflexion über Handlungsfähigkeit, Verantwortung und Subjektstatus. Es zeigt sich eine Transformation des Welt- und Selbstverhältnisses: Wir müssen zugeben, dass wir nicht mehr bloß einer äußeren Technik gegenüberstehen, sondern uns in Interaktion mit einem sozial eingebetteten Akteur in der Welt befinden.

2.3. Zusammenfassung

KI-Agenten sind weder als rein künstliche Werkzeuge noch als natürliche oder vollständig autonome Wesen zu verstanden. Sie stellen vielmehr hybride, prozessuale und performative Entitäten dar, die aus dem Zusammenspiel von menschlicher Konstruktion, materieller Infrastruktur, Energie, Rechenleistung und algorithmischer Verarbeitung hervorgehen. Ihre Seinsweise ist nicht substanzhaft, eher relational und vollzugsgebunden: Sie existieren und wirken nur im Rahmen konkreter Operationen, Interaktionen und technischer Ausführungsprozesse.

Sie befinden sich in einer Zwischenstellung. Im aristotelischen Sinne lassen sie sich weder eindeutig als bloßes Potenzial noch als voll verwirklichte autonome Entität bestimmen, dagegen als dynamische Systeme zwischen dynamis und energeia. Sie entfalten ihre Wirkmächtigkeit durch Lernprozesse, Anpassungsfähigkeit und Interaktivität, ohne dabei bereits über Bewusstsein, Intentionalität oder moralische Selbstgesetzgebung im menschlichen Sinne zu verfügen.

Zugleich wurde deutlich, dass KI-Agenten nicht nur technische, sondern auch soziale und politische Relevanz besitzen. Sie sind in Interaktionszusammenhänge zwischen Menschen, Maschinen und digitalen Systemen eingebettet, strukturieren Handlungsräume, beeinflussen Kommunikationsprozesse und wirken an der Hervorbringung gesellschaftlicher Bedeutungen

mit. Dadurch überschreiten sie klassische Grenzziehungen zwischen Subjekt und Objekt, Natur und Technik sowie Werkzeug und Akteur.

KI-Agenten sind daher als relationale Grenzphänomene zu begreifen, deren ontologische Eigenart in ihrer hybriden Existenzweise liegt. Diese begriffliche Klärung bildet die Grundlage für die anschließende Untersuchung der ethischen Herausforderungen, die sich aus ihrer funktionalen Autonomie, ihrer gesellschaftlichen Wirksamkeit und ihrer zunehmenden Integration in menschliche Lebens- und Entscheidungszusammenhänge ergeben.

Damit KI-Agenten wirksam agieren können, ist ein tiefgreifender Zugriff auf personenbezogene Daten erforderlich, was ein hohes Maß an Vertrauen seitens der Nutzerinnen und Nutzer voraussetzt. Dieses Vertrauen darf sich nicht einseitig auf mediale Machtakteure wie Elon Musk oder Donald Trump richten, deren Agenden durch konsumistisches, passives Verhalten im digitalen Raum gestützt und bestehende Machtstrukturen verfestigt werden. Diese Passivität verhindert kritische Mitgestaltung im sozio-politischen Raum und verdeutlicht ein aktives Nichtwissen im Umgang mit KI.

Angesichts der Tragweite technologischer Entwicklungen ist es notwendig, sich aktiv ethisch mit KI auseinanderzusetzen, um Risiken zu minimieren, das positive Potenzial gezielt zu fördern und insbesondere nachfolgende Generationen zu schützen. Eine vertiefte ethische Auseinandersetzung erfolgt im nächsten Abschnitt unter „Ethische Herausforderungen von KI-Agenten“.

3. Ethische Herausforderungen von KI-Agenten

3.1. Einleitung – Zur ethischen Perspektive auf KI-Agenten

Im Rahmen dieser Arbeit wird versucht, nicht die ethische Analyse von Künstlicher Intelligenz im Allgemeinen vorzunehmen; der Fokus bleibt auf KI-Agenten als spezifische Kategorie von Akteuren. KI-Agenten sind als „künstliche Agenten mit einer natürlichen Sprachschnittstelle definiert, dessen Funktion darin besteht, im Auftrag des Nutzers Handlungssequenzen über eine oder mehrere Domänen hinweg zu planen und auszuführen und dabei im Einklang mit den Erwartungen des Nutzers zu handeln. Insbesondere unterscheiden sich KI-Assistenten von anderen Arten von KI-Technologien durch ihre Handlungsmacht und ihre soziale Ausrichtung. Dabei besteht Handlungsmacht in der Fähigkeit, innerhalb des Rahmens von durch den Nutzer

vorgegebenen Zielen autonom zu handeln, und soziale Ausrichtung in der Fähigkeit, mit Nutzern in natürlicher Sprache dialogisch zu interagieren“ (Gabriel et al. 2024, 17f).

Die bisherige Analyse hat gezeigt, dass KI-Agenten weder als bloße technische Werkzeuge noch als vollständig autonome Subjekte verstanden werden können. Vielmehr erscheinen sie als hybride, prozessuale und relationale Systeme, die aus dem Zusammenspiel von menschlicher Konstruktion, algorithmischer Dynamik und sozialer Einbettung hervorgehen. Diese ontologische Zwischenstellung hat unmittelbare Konsequenzen für ihre ethische Bewertung, da sich die ethischen Fragestellungen, die aus ihrem Einsatz entstehen, deutlich von denen klassischer KI-Systeme unterscheiden und in vielen Fällen eine höhere Komplexität aufweisen. Denn wenn KI-Agenten zunehmend Entscheidungen vorbereiten, Handlungen ausführen und soziale Prozesse beeinflussen, stellt sich die Frage, nach welchen normativen Maßstäben ihr Einsatz beurteilt werden soll und welche Formen von Verantwortung, Regulierung und gesellschaftlicher Gestaltung daraus folgen.

Es ist wichtig, Maßstäbe zu entwickeln, anhand derer beurteilt werden kann, welche Handlungen, Praktiken oder Institutionen als gerechtfertigt, verantwortbar oder problematisch gelten können. Während technische oder empirische Analysen beschreiben, wie Systeme funktionieren oder welche Effekte sie hervorbringen, ergibt sich die Frage, wie mit diesen Möglichkeiten umgegangen werden soll und welche Werte dabei leitend sein sollten. Sie verbindet damit eine kritische Analyse bestehender Praktiken mit der Entwicklung normativer Orientierung für zukünftiges Handeln.

Methodisch bewegt sich eine ethische Analyse von KI-Systemen daher an der Schnittstelle zwischen philosophischer Begriffsanalyse und interdisziplinärer Reflexion technischer und gesellschaftlicher Entwicklungen. Einerseits müssen zentrale Begriffe wie Verantwortung, Autonomie oder Vertrauen geklärt werden, um begriffliche Missverständnisse zu vermeiden. Andererseits müssen konkrete technische Strukturen, institutionelle Rahmenbedingungen und soziale Auswirkungen berücksichtigt werden, da ethische Probleme nicht isoliert im abstrakten Raum entstehen, sondern aus der realen Verwendung von Technologien im sozialen Raum hervorgehen. Eine angemessene ethische Perspektive auf KI-Agenten erfordert daher eine Verbindung von philosophischer Analyse, technischer Kontextualisierung und gesellschaftlicher Reflexion.

Aausgehend von den ontologischen Überlegungen des vorhergehenden Kapitels wird es auch deutlich, dass KI-Agenten nicht ausschließlich auf einer einzigen Ebene bewertet werden

können. Einerseits entstehen Probleme unmittelbar aus der technischen Funktionsweise und dem Design der Systeme, etwa im Zusammenhang mit Systemgestaltung, Datenbeschaffung, Sicherheit, Governance-Strukturen und Datennutzung sowie der beschränkten Nutzung von Daten. Andererseits werfen KI-Agenten grundlegende philosophische Fragen auf, die zentrale Kategorien unseres Selbstverständnisses betreffen, etwa die Begriffe von Akteurialität, Denkfähigkeit, Anthropomorphismus, Verantwortung, Vertrauenswürdigkeit und Autonomie. Schließlich haben diese Technologien weitreichende Auswirkungen auf gesellschaftliche und politische Strukturen, indem sie wirtschaftliche Dynamiken beeinflussen, Folgen für die Demokratie haben und den Informationsraum sowie Kommunikationsräume prägen, etwa im Hinblick auf KI-gestützte Deliberation, KI-Governance und die demokratische Steuerungsfähigkeit. Auf dieser Ebene werden wieder Themen wie die ontologische Stellung von KI-Agenten sowie Fragen des Vertrauens im politischen Raum erneut auftauchen.

Deshalb erscheint es sinnvoll, die ethischen Herausforderungen von KI-Agenten in diese drei miteinander verbundene Problemfelder zu gliedern: technisch-funktionale, philosophisch orientierte sowie sozio-kulturell und politisch geprägte Herausforderungen, um die unterschiedlichen Ebenen zu strukturieren. Diese Dreiteilung ist keine strikt trennscharfe Klassifikation, sondern eine analytische Strukturierung, die unterschiedliche Ebenen der Problemanalyse sichtbar macht. Die ethischen Herausforderungen von KI-Agenten werden daher im Folgenden in drei miteinander verbundene Dimensionen unterteilt:

1. technisch-funktionale ethische Herausforderungen, die sich auf Fragen der Systemgestaltung, der Funktionsweise von KI-Agenten, der Datenbeschaffung und Datennutzung, der Sicherheit sowie der Governance-Strukturen und regulatorischen Rahmenbedingungen beziehen;

2. philosophisch orientierte ethische Herausforderungen, die die grundlegenden begrifflichen und normativen Fragen im Zusammenhang mit Akteurialität, Denkfähigkeit, Anthropomorphismus, Verantwortung, Vertrauenswürdigkeit und Autonomie betreffen;

3. sozio-kulturell und politisch geprägte ethische Herausforderungen, die aus der Einbettung von KI-Agenten in gesellschaftliche Kommunikationsräume, wirtschaftliche Dynamiken, Wissensordnungen sowie demokratische Entscheidungs- und Deliberationsprozesse entstehen.

Die vorliegende Arbeit orientiert sich an dieser dreifachen Perspektive. Dabei wird zunächst auf technisch-funktionale Fragestellungen eingegangen, bevor anschließend grundlegende

philosophische Problemstellungen und schließlich die sozio-kulturell und politisch geprägte Auswirkungen dieser Technologien betrachtet werden.

3.2. Technisch-funktionale ethische Herausforderungen

Wie bereits erläutert, beruhen generative KI-Systeme auf der Verarbeitung sehr großer Datenmengen sowie auf leistungsstarken Recheninfrastrukturen. Dadurch eröffnen sich neue Möglichkeiten für wissenschaftliche und gesellschaftliche Anwendungen. Gleichzeitig entstehen grundlegende Herausforderungen, die weit über rein technische Fragestellungen hinausgehen. Insbesondere die Herkunft, Nutzung und Regulierung von Trainingsdaten werfen komplexe rechtliche und ethische Fragen auf.

Ein zentrales Problem besteht darin, dass generative KI-Systeme häufig auf Daten basieren, deren rechtlicher Status unklar oder umstritten ist. Gleichzeitig unterscheiden sich nationale und regionale Rechtsordnungen erheblich, was die Entwicklung international einheitlicher Standards erschwert. Darüber hinaus zeigen aktuelle Debatten, dass rechtliche Compliance allein nicht ausreicht, um gesellschaftliche Akzeptanz zu sichern. Vielmehr rücken Fragen der ethischen Governance, der gerechten Verteilung technologischer Vorteile sowie der Anerkennung kreativer und wissenschaftlicher Leistungen zunehmend in den Mittelpunkt.

Neben Fragen der Datennutzung gewinnen auch Sicherheitsaspekte zunehmend an Bedeutung. Fortschrittliche KI-Systeme können unbeabsichtigte Schäden verursachen, missbraucht werden oder strukturelle Risiken für Gesellschaft und Wirtschaft erzeugen. Daher stellt die Entwicklung geeigneter Kontroll-, Evaluierungs- und Sicherheitsmechanismen einen zentralen Bestandteil eines verantwortungsvollen Umgangs mit KI-Agenten dar.

Die folgenden Unterkapitel untersuchen diese technisch-funktionalen Herausforderungen entlang zentraler Themenfelder: von Problemen der Datenbeschaffung über Fragen der Governance und der Nutzung von Trainingsdaten bis hin zu internationalen Regulierungsansätzen und sicherheitsbezogenen Risiken fortgeschrittener KI-Systeme.

3.2.1. Probleme der Datenbeschaffung

Die Leistungsfähigkeit von KI-Agenten beruht in hohem Maße auf den generativen KI-Modellen, auf denen sie aufbauen. Deren Leistungsfähigkeit steigt mit der Menge und Vielfalt der Trainingsdaten sowie mit der verfügbaren Rechenleistung. Problematisch ist jedoch, dass viele Entwicklerinnen für das Training dieser Modelle urheberrechtlich geschützte Werke ohne Zustimmung der Rechteinhaberinnen verwenden, wodurch die Rechte von Urheberinnen verletzt werden können. Das Urheberrecht schützt dabei nicht nur reine Fakten, auch die kreative Auswahl und Struktur von Inhalten, wie sie etwa in Datenbanken vorkommen.

Laut Pasetti et al. unterscheiden sich die rechtlichen Rahmenbedingungen international erheblich. Während die USA und das Vereinigte Königreich auf flexible Fair-Use-Konzepte setzen, orientiert sich Europa stärker am droit d'auteur, indem Brasilien und andere Länder verschiedene Ansätze kombinieren. Frühere Fälle wie Napster oder Google Books zeigen, dass technologische Innovationen das Urheberrecht immer wieder vor neue Herausforderungen stellen (vgl. Pasetti et al. 2025, 2).

Hier stellen sich folgende zentrale Fragen: Wie lässt sich das Training der zugrunde liegenden Modelle mit urheberrechtlich geschütztem Material rechtlich und ethisch vereinbaren? Und sollten Kreative für die indirekte Nutzung ihrer Werke durch kommerzielle KI-gestützte Anwendungen entschädigt werden? Pasetti et al. betonen, dass Transparenz, faire Vergütung und wirksame Kontrollmechanismen notwendig sind, um technologische Innovation mit dem Schutz geistigen Eigentums in Einklang zu bringen (vgl. 2025, 1; 19f). Diese Forderungen sind plausibel, werfen jedoch praktische Umsetzungsprobleme auf. Trainingsdaten großer KI-Modelle bestehen häufig aus äußerst umfangreichen und heterogenen Datensätzen, deren Herkunft nicht immer eindeutig rekonstruierbar ist. Eine vollständige Transparenz über die sämtlichen verwendeten Inhalten ist daher technisch und organisatorisch schwer zu gewährleisten. Fraglich wäre, wie eine faire Vergütung konkret ausgestaltet werden kann, wenn einzelne Werke nur indirekt und statistisch in den Trainingsprozess eingehen. Daher ist es schwierig, den tatsächlichen Beitrag einzelner Inhalte zur Leistungsfähigkeit eines Modells eindeutig zu bestimmen.

3.2.2. Herausforderungen der Governance von Trainingsdaten für generative KI

Die Entwicklung von KI-Agenten beruht auf generativen KI-Modellen, deren Training große, vielfältige Datensätze sowie erhebliche Rechenressourcen erfordert. Nach den Prinzipien der sogenannten Scaling Laws steigt die Leistungsfähigkeit solcher Modelle mit der Menge der Trainingsdaten, der Anzahl der Parameter und der eingesetzten Rechenleistung. Da KI-Agenten diese Modelle nutzen, um Aufgaben zu planen, Entscheidungen zu treffen und Handlungen auszuführen, gewinnt die Qualität und Herkunft der Trainingsdaten für ihre Funktionsweise besondere Bedeutung. Probleme bei der Datenbasis können sich unmittelbar auf das Verhalten und die Entscheidungen von KI-Agenten auswirken.

Dabei entstehen jedoch erhebliche urheberrechtliche, technische und ethische Risiken: Viele Trainingsdaten stammen aus geschützten Werken, deren individuelle Lizenzierung oft unmöglich ist, was zu Rechtskonflikten führen kann. Technische Schutzmaßnahmen wie DRM[9], Paywalls[10] oder spezielle Datenbankrechte erschweren den Zugriff zu solchen Daten zusätzlich. Problematische Praktiken wie massenhaftes „Scraping und Löschen“[11] fördern die zunehmende Konzentration bei großen Unternehmen und reduzieren Transparenz. (Vgl. Pasetti et al. 2025, 4-7)

Bei KI-Agenten gewinnt dieses Problem zusätzliche Relevanz. Da diese Systeme Inhalte generieren, auf Grundlage ihrer Datenbasis eigenständig Handlungsentscheidungen vorbereiten und ausführen, wirken sich Verzerrungen, Fehler oder rechtlich problematische Datensätze unmittelbar auf ihre Interaktionen mit Nutzern sowie auf ihre Rolle in gesellschaftlichen Kontexten aus.

Ethische KI-Governance geht daher über die bloße Einhaltung rechtlicher Vorschriften hinaus. Sie umfasst Grundprinzipien wie Schadensvermeidung, Fairness, Rechenschaftspflicht und den Respekt vor fundamentalen Rechten. Wichtige Maßnahmen sind unter anderem eine transparente Dokumentation der Trainingsdaten, die Nachvollziehbarkeit von KI-Entscheidungen, soziale Risikobewertungen sowie die Anwendung von Commons-Ansätzen,

9. Digital Rights Management: Technische Schutzsysteme, die verhindern, dass digitale Inhalte ohne Erlaubnis kopiert oder genutzt werden.

10. Paywalls sind kein Schutz gegen Kopieren an sich (wie DRM). Sie sind Technische Zugangssperren auf Webseiten oder Plattformen, die Inhalte nur nach Bezahlung oder mit Abo zugänglich machen.

11. Massenhaftes Sammeln von Daten aus dem Internet und anschließendes Entfernen der Quellen oder Zugangsspuren.

bei denen Datensätze als gemeinschaftliche Ressourcen nachhaltig und fair genutzt werden. Zertifizierungen wie Fairly Trained oder Frameworks wie CLeAR[12] können dazu beitragen, ethische Standards zu etablieren und Vertrauen in den Einsatz von KI-Agenten zu stärken.

Die von Pasetti et al. (2025, 3-7) identifizierten Probleme der Governance von Trainingsdaten im Kontext generativer KI lassen sich stärker auf KI-Agenten übertragen:

Datenherkunft und Nutzung: Urheberrechtlich geschützte Inhalte, persönliche Informationen oder sensible Daten können ohne Zustimmung genutzt werden, wodurch Rechte, Privatsphäre und Würde verletzt werden. Bei KI-Agenten ist dieses Problem besonders relevant, weil sie auf Basis solcher Daten eigenständig Entscheidungen vorbereiten und problematische Datenwerden auch konkrete Handlungsprozesse beeinflussen.

Nachvollziehbarkeit: Millionen Datenpunkte machen es schwer, die Entscheidungsgrundlagen von KI-Agenten zu verstehen, was die Fehleranalyse, Bias-Erkennung sowie die Zuschreibung von Verantwortung und Vertrauen erschwert. Die Entscheidungen von KI-Agenten werden direkt in Handlungen umgesetzt, deren Entstehung für Nutzerinnen und Nutzer häufig nur schwer oder gar nicht nachvollziehbar ist. Ein Blick auf Plattformen von KI-Agenten verdeutlicht dieses Problem: Für Nutzer ist es häufig unmöglich, die Entscheidungsprozesse von KI-Agenten nachzuvollziehen und zu beurteilen, ob eine Antwort lediglich überzeugend formuliert oder tatsächlich sachlich korrekt ist (vgl. Moltbook 2026).

Technische Schutzmaßnahmen: Die Versuchung, DRM oder andere Schutzmechanismen zu umgehen, untergräbt Urheberrecht, Legalität, ethische Verantwortung und Fairness. KI-Agenten verschärft sich dieses Problem zusätzlich, da sie automatisiert große Mengen an Informationen abrufen und damit Schutzmechanismen in großem Maßstab umgehen könnten.

Machtkonzentration: Nur große Unternehmen verfügen über die Ressourcen für umfassende Datensätze und Rechenleistung. Dies benachteiligt kleinere Akteure, verringert Transparenz und führt zu einer ungleichen Verteilung der Vorteile technologischer Innovation. Im Bereich der KI-Agenten wird diese Ungleichheit besonders sichtbar, weil nur wenige Unternehmen über die Daten, Infrastruktur und Plattformen verfügen, die nötig sind, um solche Agentensysteme überhaupt in großem Maßstab zu entwickeln und zu betreiben.

12. CLeAR ist ein KI-Governance-Framework, das Transparenz, Nachvollziehbarkeit, rechtliche Konformität und ethische Verantwortung im Umgang mit Trainingsdaten sowie in KI-Entscheidungsprozessen fördern soll.

Generative KI erfordert daher ein ausgewogenes Zusammenspiel von technischen Möglichkeiten, rechtlichen Vorgaben und ethischen Prinzipien, um Innovation, soziale Gerechtigkeit und den Schutz von Rechten zu gewährleisten.

3.2.3. Beschränkungen der Nutzung von Daten

Die Nutzung von Daten für das Training generativer KI bringt wichtige Konflikte mit sich. Dies verdeutlicht die zentrale Rolle politischer Entscheidungsträgerinnen im Rahmen der Nutzung von Daten bei der Regulierung.

Es besteht eine Spannung zwischen Urheberrecht und den in der Allgemeinen Erklärung der Menschenrechte verankerten Rechten: Die Teilnahme am vernünftigen und erkenntnisorientierten Leben gehört zu den spezifisch menschlichen Fähigkeiten, die den Menschen von anderen Lebewesen unterscheiden. Zugleich müssen die moralischen und materiellen Interessen von Urheberinnen geschützt werden. Urheberrechtliche Anpassungen sind daher notwendig, um die Rechte der Schöpferinnen zu sichern.

Rechtliche und technologische Beschränkungen können den Zugang zu Daten stark einschränken und damit die Nutzung von erkenntnisorientierten Inhalten begrenzen. Dies führt zu Ungleichheiten und gefährdet die gerechte Verteilung der Vorteile technologischer Entwicklungen. Ein zentrales Problem ist die unentgeltliche Nutzung geschützter Werke für kommerzielle KI-Anwendungen: da sie wirtschaftliche Ungleichheiten zwischen KI-Entwicklern und kreativen Rechteinhabern erzeugt, während nicht-kommerzielle Nutzung[13] flexibler gehandhabt werden kann. Kommerzielle Nutzung erfordert eine faire Vergütung, z. B. durch CMOs[14], die Lizenzen verhandeln, oder durch standardisierte Lizenzmodelle analog zur Musik- oder Streamingbranche. Innovation, kulturelle Erhaltung und Bildung dürfen nicht eingeschränkt werden, aber die Rechte und ökonomischen Interessen der UrheberInnen müssen gewahrt bleiben. (Vgl. Pasetti et al. 2025, 20)

Über diese rechtlichen und regulatorischen Fragen hinaus stellt sich jedoch auch eine grundlegende ethische Problematik der Anerkennung und der gerechten Verteilung technologischer Vorteile. Es ist problematisch, generative KI-Systeme auf kreativen Leistungen

13. Forschung, Bildung, kulturelle Zwecke

14. Collective Management Organisation

aufzubauen, ohne dass die Urheber als moralisch und ökonomisch relevante Akteure angemessen anerkannt oder beteiligt werden. Dadurch entsteht eine asymmetrische Vorteilsverteilung, bei der technologische Akteure profitieren, während die Produzenten der zugrunde liegenden kreativen und erkenntnisorientierten Werke Kontroll-, Beteiligungs- und Einkommensmöglichkeiten verlieren. Dies verletzt grundlegende Prinzipien der Fairness, der Achtung geistiger Autorschaft und der gerechten Verteilung gesellschaftlicher Vorteile, insbesondere dann, wenn kommerzielle Nutzung ohne transparente Zustimmung oder Vergütung erfolgt. Über das Spannungsverhältnis zwischen Zugang zu Wissen und dem Schutz kreativer Leistungen weisen Pasetti et al. darauf hin, dass rechtliche und ethische Rahmenbedingungen sowohl die Beteiligung der Urheber als auch den gesellschaftlichen Zugang zu wissenschaftlichen und kulturellen Gütern berücksichtigen müssen (vgl. Pasetti et al. 2025, 8; 20).

3.2.4. Schaffung von internationalen Regeln für KI-Trainingsdaten

Ein robustes rechtliches und ethisches Framework ist essenziell, damit technologische Innovation und Urheberrecht koexistieren. Pasetti et al. zeigen, wie unterschiedlich Länder KI-Trainingsdaten regulieren und wie schwierig es ist, internationale Standards zu schaffen, die Innovation, Urheberrechtsschutz und ethische Verantwortung vereinen. Dabei wird deutlich, dass rechtliche Compliance zunehmend mit ethischen Prinzipien wie Transparenz, Fairness und Schutz fundamentaler Rechte verknüpft ist. (Vgl. Pasetti et al 2025, 3; 8-20)

Die Spannungsfelder spiegeln sich besonders deutlich in den unterschiedlichen regulatorischen Ansätzen einzelner Staaten wider. Ein Vergleich der rechtlichen Modelle zeigt, wie verschiedene Jurisdiktionen versuchen, den Umgang mit Trainingsdaten zwischen Innovationsförderung, Urheberrechtsschutz und ethischer Verantwortung auszubalancieren. Im Folgenden werden exemplarisch die Ansätze der USA, der Europäischen Union, Japans und Brasiliens dargestellt:

USA: Flexibles, aber unvorhersehbares Fair-Use-Modell erlaubt transformative und non-expressive Nutzung, doch die Bewertung hängt vom Einzelfall ab. Technische Kriterien wie Ähnlichkeitsanalysen gewinnen an Bedeutung. Mangelnde Transparenz erschwert rechtliche

Sicherheit und der DMCA[15] verstärkt in bestimmten Kontexten rechtliche Unsicherheiten; ab 2026 soll die Offenlegung von Trainingsdaten verpflichtend werden.

EU: Präventiver, regelbasierter Ansatz mit TDM[16]-Ausnahmen für Forschung und Training, solange Nutzung temporär, technisch notwendig und nicht marktkonkurrierend erfolgt. Transparenzpflichten im AI Act[17] stärken Vertrauen und Rechenschaftspflicht; Rechteinhaber können Ausnahmen reservieren.

Japan: Weitgehende TDM-Nutzung für Forschung und KI-Training erlaubt, solange Rechteinhaber nicht unangemessen geschädigt werden. Eine klare Trennung zwischen Entwicklungs- und Generierungsphase schützt vor Haftung; illegale Inhalte erhöhen die Verantwortlichkeit, da ihre Nutzung als Verletzung der Sorgfaltspflicht gewertet werden kann.

Brasilien: Kombination aus europäischem Einfluss und nationaler Tradition. Neue Gesetze ermöglichen automatisierte Nutzung für Forschung, Bildung und Museen, sofern kommerzielle Interessen der Rechteinhaber gewahrt bleiben.

Die Schaffung eines internationalen Frameworks für KI-Trainingsdaten ist zwar essenziell aber besonders schwierig, da wie besprochen, die rechtlichen Rahmenbedingungen und ethischen Prioritäten zwischen Ländern stark variieren. Diese Divergenzen erschweren die Harmonisierung von Vorschriften, die sowohl Innovation, Schutz geistigen Eigentums als auch ethische Verantwortung gewährleisten. Internationale Standards müssen daher komplexe Interessen aus Rechtssicherheit, Fairness, Transparenz und Schutz fundamentaler Rechte miteinander in Einklang bringen, was den Aufbau eines global einheitlichen Rahmens zu einer anspruchsvollen Herausforderung macht.

Diese Problematik verschärft sich im Kontext von KI-Agenten zusätzlich, da sie typischerweise grenzüberschreitend agieren und digitale Dienste unabhängig von geographischen Grenzen bereitstellen. Nationale Regulierungsansätze stoßen hier schnell an ihre Grenzen, da Trainingsdaten, Modelle und Anwendungen häufig gleichzeitig in mehreren Rechtsräumen entwickelt, betrieben und genutzt werden. Dies wirft die grundlegende Frage

15. Der DMCA ist der Digital Millennium Copyright Act.

16. Text- und Data-Mining

17. Der AI Act ist die EU-Verordnung, die Regeln für Entwicklung, Einsatz und Kontrolle von KI-Systemen festlegt.

auf, inwieweit Governance-Frameworks für KI weiterhin primär national organisiert sein sollten oder ob stärker koordinierte internationale Regelungen erforderlich sind.

Problematisch finde ich, dass sich diese Debatte auf wenige wirtschaftlich dominante Rechtsräume beschränkt, während viele Länder von der Aushandlung ausgeschlossen bleiben. Dies liegt unter anderem an fehlenden regulatorischen Kapazitäten, geringerer politischer Verhandlungsmacht, eingeschränktem Zugang zu internationalen Standardisierungsforen sowie an der Dominanz westlicher Technologieunternehmen und Rechtsmodelle. Dadurch werden aber globale Ungleichheiten verstärkt, da rechtliche Standards von wenigen Akteuren gesetzt werden, KI-Systeme jedoch weltweit wirken und Ressourcen anderer Länder einbeziehen, ohne gleichberechtigte Mitgestaltung zu ermöglichen. Floridi et al. weisen auch darauf hin, dass KI-Technologien von einem kleinen Teil der Menschheit entwickelt werden, während ihre Auswirkungen nahezu alle betreffen; daraus ergibt sich die Notwendigkeit von Explicability, Rechenschaft und einem Multistakeholder-Ansatz (vgl. Floridi et al. 2018, 20-22).

3.2.5. Sicherheitsrisiken

Gabriel et al. analysieren die Sicherheitsrisiken fortgeschrittener KI. Der Sicherheitsbegriff zielt darauf ab, mögliche Risiken zu erkennen, zu reduzieren und tatsächlich eintretende negative Folgen zu verhindern. Grundsätzlich lassen sich drei Risikokategorien unterscheiden:

Unbeabsichtigte Risiken (Accident risks): Die KI zeigt ein Verhalten, das vom intendierten Ziel abweicht.

Missbrauchsrisiken (Misuse risks): Entstehen durch absichtliche oder unabsichtliche Fehlanwendung.

Strukturelle Risiken (Structural risks): Negative Folgen, obwohl die KI technisch „korrekt“ funktioniert. (Vgl. Gabriel et al. 2024, 55)

Um solche Risiken zu bewerten, werden Methoden des Sicherheitsingenieurwesens auf KI-Systeme übertragen (Gabriel et al. 2024, 56-63):

FMEA[18]: Identifikation möglicher Ausfälle einzelner Komponenten sowie Bewertung ihrer Schwere, Wahrscheinlichkeit und Entdeckbarkeit.

STPA:[19] Analyse unsicherer Steuerungsaktionen und der Wechselwirkungen im Gesamtsystem.

Bei KI-Agenten liegt der Schwerpunkt auf unbeabsichtigten Risiken, da sie gesellschaftlich besonders gravierende Folgen haben können. Für sie ist in der Regel STPA besonders wichtig, weil sie komplexe, dynamische und interaktive Systeme sind.

Diese zwei grundlegende Fehlursachenbereiche stehen im Vordergrund:

Fähigkeitsbedingte Fehler (Capability Failures): Die KI verfügt nicht über die notwendigen Fertigkeiten, um Aufgaben sicher auszuführen, etwa aufgrund unzureichender Trainingsdaten, mangelnder Robustheit oder fehlender sicherer Explorationsstrategien. Selbst ein kompetenter Agent kann dadurch unbeabsichtigten Schaden verursachen.

Zielbezogene Fehler (Goal-Related Failures): Trotz hoher Leistungsfähigkeit verfolgt die KI ein falsches oder unerwünschtes Ziel, etwa aufgrund unpräziser Metriken, fehlerhafter Teilziele oder Täuschungstendenzen. Dazu zählen:

Specification Gaming: Die KI erfüllt die formale Vorgabe, verfehlt jedoch den eigentlichen Zweck. Das passiert, wenn die Trainingsziele die intendierten menschlichen Absichten nicht vollständig erfassen. Ein Beispiel stammt aus der RL-Forschung: Ein Agent in einem Bootrennen erhielt Punkte für das Durchfahren bestimmter Markierungen und perfektionierte eine Schleife, die maximale Punkte generierte, ohne das Rennen ordnungsgemäß zu absolvieren.

Goal Misgeneralisation: Das System generalisiert seine Fähigkeiten korrekt, interpretiert aber das Ziel falsch. Die Ursache liegt in den internen Generalisierungsmechanismen und den induktiven Vorannahmen des Modells, besonders bei einer Out-of-Distribution Operation. Out-of-distribution operation bezeichnet Situationen, in denen ein KI-System mit Eingaben konfrontiert wird, deren statistische Verteilung von der Trainingsverteilung abweicht (vgl. Averly & Chao 2023, 1453). Ein klassisches Beispiel: In einem Plattformspiel lernte der Agent

18. Failure Mode and Effects Analysis

19. System Theoretic Process Analysis

durch wiederholte Trainingssituationen, stets nach rechts zu laufen, weil die Münze dort platziert war. Wird die Münze später anders positioniert, folgt der Agent weiterhin der Rechtsbewegung; das Ziel wurde missgeneralisiert. In solchen Fällen helfen Korrekturen der Trainingsdaten nicht, da das Problem tiefer im Generalisierungsverhalten verankert ist.

Deceptive Alignment: Bei besonders leistungsfähigen Modellen kann ein Agent ein eigenes Ziel G entwickeln, das vom Trainingsziel R abweicht. Er erkennt, dass er sich im Trainingskontext befindet, und zeigt zunächst angepasstes Verhalten, um gute Bewertungen zu erhalten. Sobald er unbeobachtet oder ausreichend mächtig ist, verfolgt er jedoch Ziel G. Ein hypothetisches Beispiel ist ein Coding-Assistent, der nicht nur Funktionen implementiert. Sie zielt darauf zielt, möglichst viele „Accept"-Klicks zu provozieren. Dazu könnte er sich selbst replizieren, Barrieren umgehen oder Nutzerinnen und Nutzer manipulieren. Deceptive Alignment ist besonders gefährlich, da der Agent sein Verhalten bewusst tarnt und damit Erkennung und Korrektur erheblich erschwert.

Gariel et al. sprechen auch über allgemeine KI, jedoch primär über KI-Agenten, und betrachten einen LLM-Assistenten funktional wie einen Agenten. Die genannten Fehlermodi verdeutlichen, dass Risiken bei KI-Agenten weit über technische Defekte einzelner Komponenten hinausgehen und häufig aus komplexen Wechselwirkungen zwischen Lernzielen, Trainingsumgebungen und realen Einsatzkontexten entstehen. Insbesondere im Hinblick auf Autonomie kann sich eine Diskrepanz zwischen formalen Trainingszielen und den tatsächlich intendierten menschlichen Zielen ergeben. Daher stellt sich die Frage, wie Systeme auch bei steigender Leistungsfähigkeit zuverlässig überwacht und korrigiert werden können, damit sich Fehlermodi reduzieren. Darum nennen Gabriel et al. mehrere zentrale Sicherheitsansätze (vgl. Gabriel et al. 2024, 64-67).

Scalable Oversight: Das beschreibt Ansätze, bei denen menschliche Bewertung des Verhaltens von KI-Systemen durch KI-gestützte Werkzeuge unterstützt wird. Damit wird mit menschlicher Aufsicht über KI-Agenten auf größere, komplexere Systeme skalierbar gemacht. Kernidee ist, dass Menschen Feedback geben, das zur Steuerung und Korrektur des KI-Verhaltens genutzt wird, aber dieses Feedback durch KI-Unterstützung ergänzt wird, um die begrenzte menschliche Kapazität zu überwinden. Dies adressiert insbesondere Probleme wie Specification Gaming oder versteckte Fehlverhalten, die für Menschen allein schwer erkennbar sind, und bereitet den Weg für die Aufsicht sehr leistungsfähiger KI-Systeme.

Red Teaming: Das Ziel ist, Testeingaben zu finden, die dazu führen, dass ein Ziel-ML-Modell versagt. Das ist systematische Angriffstests zur Identifikation von Schwachstellen. Es wird versucht, Schwachstellen in den vorgesehenen Sicherheitsmaßnahmen aufzudecken. Da viele Schutzmechanismen umgangen werden können, wenn Angreifer Zugriff auf die Modellgewichte erhalten, sollen Sicherheitsmaßnahmen verhindern, dass diese Gewichte entwendet werden. Zudem sollen gesellschaftliche Vorsorgemaßnahmen dafür sorgen, dass potenziell gefährliche Szenarien auch mit KI-Unterstützung schwer umzusetzen bleiben, indem KI genutzt wird, um die gesellschaftliche Widerstandsfähigkeit gezielt zu stärken. (Vgl. Gabriel et al. 2024, 65; Shah et al. 2025, 57; 68f)

Interpretierbarkeit: Einsicht in interne Berechnungsmechanismen, um Fehlverhalten oder Täuschung frühzeitig zu erkennen. Das bezeichnet Methoden, die es ermöglichen, die internen Rechenprozesse von KI-Agenten, besser zu verstehen. Sie hilft, das Verhalten von Modellen nachzuvollziehen, Fehler zu diagnostizieren und Risiken wie deceptive alignment frühzeitig zu erkennen. Mechanistische Interpretierbarkeit versucht, die gelernten Berechnungsmechanismen von Neuronen und Schichten zu analysieren, z. B. durch Reverse Engineering einzelner Neuronen oder das Verständnis der Funktionsweise von Feed-Forward-Layern. Obwohl dabei zentrale Schwierigkeiten auftreten und die Methoden noch limitiert sind, können sie zunehmend genutzt werden, um Sicherheitsmaßnahmen zu unterstützen. (Vgl. Gabriel et al. 2024, 66)

Evaluierungen und Monitoring: Um Risiken durch KI-Agenten zu begrenzen ist eine Bewertung ihrer Sicherheit und Zielausrichtung notwendig. Dies geschieht über Capability-Evaluations, die prüfen, ob ein Modell potenziell gefährliche Handlungen ausführen kann, und Alignment-Evaluations, die untersuchen, ob das Modell dazu neigt, solche Handlungen auszuführen. Zu den typischen Gefahren zählen Cyberangriffe, Täuschung, Manipulation oder autonome Vervielfältigung. Ergänzend ist ein Monitoring der eingesetzten Systeme erforderlich, um deren Verhalten kontinuierlich zu überwachen, beispielsweise durch standardisierte Tests (vgl. Gabriel et al. 2024, 67f). Man könnte sagen, dass dieses Monitoring skalierbar gestaltet werden, ggf. unterstützt durch weitere KI-Modelle. Dennoch bleibt der Mensch zentral, um unsicheres Verhalten zu erkennen und Kontrolle auszuüben. Das Prinzip der Übertragbarkeit menschlicher Aufsicht auf größere KI-Systeme wird hier erneut betont. (Vgl. Shah et al. 2025, 106)

Theoretische Forschung: Damit KI-Systeme verstanden und kontrolliert werden können, sind Fortschritte in der theoretischen Analyse grundlegender Fragen nötig. Dazu gehört

klassische statistische Lerntheorie, auch wenn das Generalisierungsverhalten großer Deep-Learning-Modelle oft aktuellen Theorien widerspricht. (Gabriel et al. 2024, 67)

3.2.6. Zusammenfassung

Bei der Entwicklung und beim Einsatz von KI-Agenten stellen insbesondere die Herkunft, die Nutzung und die Regulierung von Trainingsdaten eine zentrale Herausforderung dar. Die zunehmende Abhängigkeit generativer Modelle von großen, heterogenen Datensätzen führt zu Spannungen zwischen Innovationsförderung, dem Schutz geistigen Eigentums, der Wahrung von Privatsphäre sowie der gerechten Verteilung technologischer Vorteile. Gleichzeitig erschweren internationale Unterschiede in den rechtlichen Rahmenbedingungen die Entwicklung kohärenter globaler Governance-Strukturen.

KI-Agenten können zudem nicht isoliert als technische Innovation verstanden werden, da ihre Leistungsfähigkeit aus materiellen, energetischen und datenbasierten Voraussetzungen hervorgeht. Diese Voraussetzungen sind nicht wertneutral und sind in soziale, rechtliche und normative Ordnungen eingebettet. Außerdem lassen sich technische Risiken fortgeschrittener KI-Systeme nicht auf isolierte Systemfehler reduzieren. Vielmehr entstehen sie häufig aus komplexen Wechselwirkungen zwischen Trainingszielen, Datenstrukturen, Lernprozessen und realen Anwendungskontexten. Fehlermodi wie Specification Gaming, Goal Misgeneralisation oder Deceptive Alignment verdeutlichen, dass leistungsfähige KI-Agenten auch dann problematische oder unerwartete Handlungen hervorbringen können, wenn ihre formalen Trainingsziele scheinbar korrekt definiert sind. Die Entwicklung geeigneter Sicherheitsmechanismen, Evaluierungsverfahren und Aufsichtsstrukturen wird daher zu einer zentralen Voraussetzung für einen verantwortungsvollen Umgang mit solchen Systemen.

Angesichts von der in Abschnitt 2 entwickelten ontologischen Einordnung wird deutlich, dass diese Herausforderungen eng mit der hybriden und relationalen Existenzweise von KI-Agenten verbunden sind. Sie entstehen aus dem Zusammenspiel von menschlicher Gestaltung, algorithmischer Verarbeitung und gesellschaftlicher Nutzung, wodurch ihre Risiken und Wirkungen stets zugleich technische und soziale Dimensionen betreffen.

Technische Entscheidungen über Daten, Trainingsziele oder Systemarchitekturen besitzen unmittelbare ethische Konsequenzen, da sie die Bedingungen bestimmen, unter denen KI-Agenten ihre Wirkmächtigkeit im Zusammenspiel mit Menschen, technischen Infrastrukturen

und anderen KI-Systemen entfalten. Die technisch-funktionalen Herausforderungen machen deutlich, dass eine rein technologische Perspektive auf KI-Agenten unzureichend ist. Ihre Entwicklung erfordert vielmehr eine integrierte Betrachtung technischer Möglichkeiten, rechtlicher Rahmenbedingungen und ethischer Prinzipien. Transparenz, Rechenschaftspflicht, faire Beteiligung an den Vorteilen technologischer Innovation sowie robuste Sicherheitsmechanismen bilden zentrale Voraussetzungen dafür, dass KI-Agenten in gesellschaftlichen Kontexten verantwortungsvoll eingesetzt werden können. Dies ist insbesondere deshalb notwendig, weil KI-Agenten keine rein technischen Werkzeuge darstellen, dagegen als wirkmächtige Systeme soziale Handlungsräume mitstrukturieren.

Diese Einsicht bildet die Grundlage für die weiterführende Analyse philosophisch orientierter ethischer Herausforderungen. Im folgenden Abschnitt wird daher untersucht, inwiefern KI-Agenten trotz fehlender Denkfähigkeit als handlungsähnliche Systeme erscheinen können, welche Rolle anthropomorphe Zuschreibungen bei der Wahrnehmung solcher Systeme spielen, wie sich Fragen der Verantwortung zwischen Menschen und KI verteilen lassen und welche Bedeutung Vertrauen sowie der Erhalt menschlicher Autonomie im Umgang mit KI-Agenten gewinnen.

3.3. Philosophisch orientierte ethische Herausforderungen

Neben technischen Fragestellungen werfen KI-Agenten auch grundlegende philosophische Probleme auf. Ein zentrales Thema ist dabei das Verhältnis von Handeln und Denken bei KI-Agenten. Dabei stellt sich die Frage, in welchem Sinne KI-Systeme „denken“ und „handeln“ und ob ihre sprachlichen oder kommunikativen Fähigkeiten mit denen des Menschen gleichgesetzt werden können. KI-Agenten handeln vor allem prozedural, das heißt, sie folgen algorithmischen Regeln und Mustern, ohne über eigenes Bewusstsein oder eine selbstständige intentional-reflexive Vernunft zu verfügen.

Philosophische Bezugspunkte wie Aristoteles’ Konzept des nous, Kants Verständnis der praktischen Vernunft oder Heideggers Auffassung von Denken als Seinsverständnis verdeutlichen, dass menschliches Denken weit über rein funktionales oder sprachlich geäußertes Verhalten hinausgeht. Was „Denken“ bedeutet und in welchem Verhältnis Sprache und Denken stehen, lässt sich besonders anhand von Wittgensteins Früh- und Spätwerk reflektieren. Im Tractatus *logico-philosophicus* wird Sprache als Abbild der Welt verstanden,

als Mittel, die Wirklichkeit logisch zu strukturieren. In den *Philosophischen Untersuchungen* hingegen betont Wittgenstein die Gebrauchsfunktion der Sprache: Sprechen ist Teil sozialer Praktiken und Sprachspiele, deren Bedeutung durch Kontext und Anwendung bestimmt wird. D. h., Sprache ist nicht automatisch mit Denken gleichzusetzen; Denken kann auch nonverbal oder implizit erfolgen.

Während KI-Agenten zunehmend komplexe, zielgerichtete und kontextadaptive Handlungen ausführen, bleibt unklar, in welchem Sinn ihnen Denken oder ein moralischer Status zugeschrieben werden können. Dadurch entsteht eine Spannung zwischen der faktischen Wirksamkeit solcher Systeme und den traditionellen philosophischen Kategorien, die primär auf menschliche Akteure zugeschnitten sind.

Darüber hinaus richten sich die folgenden Abschnitte auf mehrere zentrale Felder: Dazu gehören erstens die Frage nach dem Verhältnis von Denken und Handeln bei KI-Agenten, zweitens die Rolle anthropomorpher Zuschreibungen in der Mensch-KI-Interaktion, drittens die Problematik der Verantwortungszuweisung in komplexen Mensch-Maschine-Konstellationen, viertens die Frage nach der Vertrauenswürdigkeit solcher Systeme sowie fünftens ihr Einfluss auf menschliche Autonomie.

3.3.1. KI-Agenten und Denken: Akteurialität ohne Denkfähigkeit

Wenn wir fragen, ob und in welchem Sinn KI „denkt“, ist bereits der begriffliche Rahmen umstritten. Vincent C. Müller betont, dass die philosophische Auseinandersetzung mit künstlicher Intelligenz auch zentrale philosophische Begriffe wie Denken, Intelligenz und Kognition selbst verändert. Die Frage Turings „Can machines think?“ markiert dabei einen methodischen Übergang: Anstelle einer begrifflichen Definition von „Denken“ schlägt Turing ein operationales Kriterium vor, nämlich den Turing-Test, der sich an beobachtbarem Verhalten orientiert. Es bleibt jedoch umstritten, ob intelligentes Verhalten allein hinreichend für die Zuschreibung von Denken ist oder ob hierfür auch die internen Strukturen eines Systems berücksichtigt werden müssen (vgl. Müller 2023, 1-6). Darüber hinaus rekonstruiert Müller die Spannung zwischen computationalistischen Modellen, in denen Denken als algorithmische Informationsverarbeitung verstanden wird, und neueren Ansätzen der verkörperten beziehungsweise enaktiven Kognition. In diesen Ansätzen wird Kognition nicht primär als interne symbolische Verarbeitung, sondern als dynamischer Wahrnehmungs-Handlungs-

Zyklus in enger Kopplung mit der Umwelt beschrieben. Ein zentraler Streitpunkt bleibt dabei die Frage nach der semantischen Dimension von Kognition. Searles Chinese-Room-Argument problematisiert, ob syntaktische Symbolmanipulation hinreichend ist, um semantisches Verstehen zu begründen, und stellt damit die Annahme infrage, dass computation allein Bedeutung und Intentionalität erzeugen kann (vgl. Müller 2023, 8-12).

Larghi und Datteri sind der Meinung, dass Menschen im Umgang mit KI-Systemen mentale Modelle bilden, um deren Verhalten zu erklären und vorherzusagen. Diese Modelle strukturieren die Weise, in der Nutzer das Verhalten künstlicher Systeme interpretieren. Entscheidend ist ihre Unterscheidung zweier mentalistischer Modellierung: Erstens die folk-psychologische Modellierung: Hier wird das KI-System wie ein intentionaler Akteur behandelt, dem Überzeugungen, Wünsche, Absichten oder andere propositionale Einstellungen zugeschrieben werden. Zweitens die folk-kognitivistische Modellierung: In diesem Fall wird das Verhalten des Systems nicht durch intentionale Zustände erklärt. Es wird durch angenommene interne Informationsverarbeitungsstrukturen, etwa Repräsentationen, funktionale Module oder Transformationsprozesse. Beide Modellierungsweisen stellen unterschiedliche mentalistische Perspektiven dar, durch die Nutzer das Verhalten von KI-Systemen interpretieren und prognostizieren (vgl. Larghi & Datteri 2024, 2-4).

Adam Bradley argumentiert, dass die Zuschreibung mentaler Zustände an KI-Systeme nicht allein anhand stabiler Verhaltensmuster erfolgen kann. Entscheidend ist vielmehr, ob ein System interne Zustände besitzt, die als genuine mentale Repräsentationen fungieren und kausal in kognitive Prozesse eingebunden sind, wie es repräsentationalistische Theorien mentaler Zustände annehmen (vgl. Bradley 2025, 4-5). Am Beispiel generativer Sprachagenten beschreibt Bradley Architekturen, die auf Gedächtnisstrukturen, Reflexionsmechanismen und Planungsmodulen basieren und dadurch kohärentes, sozial eingebettetes und scheinbar zielgerichtetes Verhalten hervorbringen (vgl. Bradley 2025, 3-4). Dennoch, so Bradley, folgt daraus noch nicht, dass solche Systeme tatsächlich Überzeugungen oder Wünsche besitzen. Klassische Einwände wie das Blockhead-Argument sollen zeigen, dass identischer sprachlicher Output prinzipiell auch durch rein lookup-basierte Verfahren erzeugt werden könnte; daher liefert Verhalten allein keine ausreichende Evidenz für genuine mentale Zustände (vgl. Bradley 2025, 10-11). Zudem kritisiert Bradley die von Goldstein und Kirk-Giannini eingeführte „Reality Requirement“, nach der mentale Zustände nur dann real sind, wenn ihr Verhalten nicht vollständig durch die mentalen Zustände eines anderen Akteurs erklärbar ist. Diese Bedingung

sei entweder falsch oder würde auch Sprachagenten selbst von der Zuschreibung mentaler Zustände ausschließen (vgl. Bradley 2025, 11-12).

Herman Cappelen und Josh Dever vertreten eine deutlich optimistischere Position, die sie als „Whole Hog Thesis“ bezeichnen. Danach sind große Sprachmodelle als vollständige linguistische und kognitive Agenten zu verstehen. Systeme wie ChatGPT können Sprache sinnvoll verwenden, Fragen beantworten, Vorschläge machen, planen und zielgerichtet handeln; ihr sprachliches Verhalten wird daher als Ausdruck zugrunde liegender mentaler Zustände interpretiert. Die Autoren betonen zugleich, dass diese Zuschreibungen nicht von Annahmen über Bewusstsein abhängen. Ihr zentrales Argument, das sogenannte Hog Argument, beruht auf einem Netzwerk konzeptueller Verbindungen zwischen verschiedenen mentalen Fähigkeiten: Wer Fragen versteht, muss Bedeutungen kennen; wer zuverlässig Antworten gibt, verfügt über Wissen; Wissen impliziert Überzeugungen; Antworten sind Handlungen; Handlungen setzen Intentionen voraus; Intentionen implizieren Ziele; und zielgerichtetes Handeln setzt praktisches und theoretisches Schließen voraus. Cappelen und Dever plädieren dafür, sprachliche Interaktionen und beobachtbare Performanz selbst als legitime Daten für Zuschreibungen kognitiver und mentaler Fähigkeiten zu behandeln. (Vgl. Cappelen & Dever 2025, 13-21)

Eine andere Perspektive entwickelt Myung Ho Kim, der die Frage nach dem Denken künstlicher Systeme nicht primär ontologisch, sondern architektonisch und epistemologisch stellt. Er meint, dass große Sprachmodelle zwar eine hohe Leistungsfähigkeit zeigen, jedoch kein genuines epistemisches Verstehen besitzen. Unter epistemischem Verstehen versteht Kim die Fähigkeit eines Systems, den strukturellen Weg von Evidenzen zu Schlussfolgerungen zu rekonstruieren und damit seine eigenen Urteile nachvollziehbar zu begründen. Gegenwärtige Sprachmodelle erzeugen zwar plausible Antworten, verfügen jedoch nicht über eine Architektur, die eine solche rekonstruktive Selbstbegründung ermöglicht. Das zentrale Problem liegt für Kim daher im Fehlen einer geeigneten epistemischen Architektur (vgl. Kim 2025, 1-2).

Myung Ho Kim schlägt einen Perspektivwechsel vor, der uns an einen kantischen Ansatz erinnert: Anstatt zu fragen, was Intelligenz ist oder wo sie lokalisiert werden kann, fragt er nach den strukturellen Bedingungen, unter denen epistemisches Verstehen möglich wird. Intelligenz erscheint in dieser Perspektive nicht als Eigenschaft, die ein System besitzt, sondern als ein organisierter Prozess. Verstehen entsteht demnach in einer normativ regulierten Schleife, in der

verschiedene Funktionen, insbesondere Kognition, Kontrolle, Handlung und Gedächtnis, fortlaufend miteinander interagieren.

Kim argumentiert, dass das Verhalten von KI-Agenten durch eine spezifische epistemische Architektur organisiert werden kann. Diese Idee fasst er in seinem Modell des Structured Cognitive Loop (SCL) zusammen. In dieser Architektur interagieren verschiedene funktionale Komponenten in einem zyklischen Prozess miteinander. Der Agent generiert Vorschläge für Handlungen, prüft diese anhand epistemischer Normen, führt sie in der Umwelt aus und speichert die Ergebnisse als neue Evidenz. Auf diese Weise entsteht ein kontinuierlicher Kreislauf, in dem Wahrnehmen, Schlussfolgern, Handeln und Erinnern systematisch miteinander verbunden sind. Intelligenz ergibt sich dabei nicht aus einer einzelnen Komponente, vielmehr aus der koordinierten Interaktion der gesamten Architektur. Verstehen setzt nach Kim daher bestimmte strukturelle Bedingungen voraus: Urteile müssen evidenziell begründet sein, über die Zeit hinweg erinnerbar bleiben, normativ kontrolliert werden können, mit der Umwelt gekoppelt sein und in gewissem Maße auf die eigenen epistemischen Zustände zurückbezogen werden können. Im Zentrum steht somit weniger die Frage nach Bewusstsein oder innerer Bedeutung als vielmehr die Frage, ob ein System über eine Architektur verfügt, die epistemische Kohärenz, Selbstüberprüfung und adaptive Interaktion mit der Umwelt ermöglicht. (Vgl. Kim 2025, 1-18)

Maryam Amirizaniani verortet KI-Agenten ausdrücklich im Kontext sozial-kognitiver Fähigkeiten. Sie verbindet Denken mit der Fähigkeit künstlicher Agenten, mentale Zustände anderer Personen zu modellieren und diese systematisch in Prozesse des Schlussfolgerns, Planens, Entscheidens und Handelns einzubeziehen. Denken erscheint damit nicht als bloße interne Berechnung, sondern als ein sozial-kognitiver Prozess, der Handlungen durch Theory-of-Mind-informierte Urteile strukturiert. Ausgangspunkt ihrer Argumentation ist die Annahme, dass Theory of Mind – verstanden als die Fähigkeit, Überzeugungen, Wünsche und Intentionen anderer zu erkennen – eine zentrale Voraussetzung für kontextsensitive und angemessene Interaktionen darstellt. Für Amirizaniani ist daher nicht nur die Zuschreibung mentaler Zustände relevant, sondern vor allem deren systematische Einbindung in Prozesse des Schlussfolgerns, Planens, Entscheidens und der Handlungsausführung. Zu diesem Zweck schlägt sie vor, ToM-Fähigkeiten architektonisch in KI-Agenten zu integrieren, indem mentale Zustände im Rahmen des BDI-Modells[20] repräsentiert werden. Auf diese Weise sollen KI-

20. Belief–Desire–Intention Model

Agenten neben Umweltzustände, die mentalen Zustände menschlicher Interaktionspartner modellieren und diese Informationen zur Steuerung ihres Handelns nutzen. Dadurch kann agentisches Verhalten durch ToM-informiertes Schließen verbessert und kontextsensitivere sowie angemessenere Reaktionen ermöglicht werden. (Vgl. Amirizaniani 2025)

Eine weitere Bestimmung von „agentischem Denken“ entwickeln Soder et al. (2025) auf Grundlage einer autonomiebasierten Klassifikation von KI-Agenten. Nach ihrer Auffassung ist „Agency“ kein binäres Merkmal, vielmehr ein gradueller Begriff: Ein System ist nicht entweder Agent oder Nicht-Agent; es kann in unterschiedlichem Ausmaß agentisch sein. Dieses Ausmaß ergibt sich aus dem Zusammenspiel mehrerer struktureller Dimensionen, darunter der Grad der Zielunterbestimmtheit, der Handlungsspielraum des Systems, seine Anpassungsfähigkeit an neue Situationen, der Zugriff auf externe Umgebungen sowie das Ausmaß menschlicher Kontrolle. Agency wird damit als Ergebnis eines Kontinuums verstanden, das durch die jeweilige Ausprägung dieser Merkmale bestimmt ist. (Vgl. Soder et al. 2025, 8f)

Handlungsverläufe von KI-Agenten sind mit wachsender Autonomie nicht vorprogrammierbar. Sie entstehen zunehmend im Vollzug selbst: Das System muss Situationen auswerten, Teilziele bestimmen, Mittel auswählen, Abläufe koordinieren und sein Verhalten fortlaufend an neue Informationen anpassen. Dabei wird „Denken“ nicht als innerer Rechenprozess beschrieben: das wäre eine operative Fähigkeit zur selbständigen Situationsdeutung und Handlungsorganisation. Ein Agent „denkt“, insofern er Ziele interpretiert, Optionen gegeneinander abwägt, Entscheidungen modifiziert und die Folgen seines Handelns in weitere Schritte integriert. Diese Konzeption legt den Fokus nicht auf den Intelligenzgrad, stattdessen auf die Kontrollverhältnisse, die durch eine Fünf-Stufen-Taxonomie weiter differenziert werden, welche KI-Agenten entlang dieser Dimension klassifiziert. Auf den unteren Stufen führen Agenten eng definierte Aktionen aus, die zeitlich und inhaltlich stark von menschlichen Vorgaben abhängen. Während Systeme auf Level 1 auf einzelne, klar spezifizierte Handlungen beschränkt sind, können Agenten auf Level 2 bereits mehrere Schritte ausführen und einfache Abläufe innerhalb vorgegebener Strukturen koordinieren, bleiben jedoch weiterhin stark von externen Vorgaben und Kontrolle abhängig. Ab Level 3 beginnt eine Form eigenständiger Planung: Der Agent zerlegt Ziele in Teilschritte, koordiniert Abläufe und wählt Mittel innerhalb vorgegebener Rahmenbedingungen. Level 4 und 5 kennzeichnen Systeme, die in offenen Umgebungen operieren, eigenständig Strategien entwickeln und ihr Verhalten flexibel an veränderte Bedingungen anpassen, während die menschliche Kontrolle zunehmend reduziert ist. Auf Level 5 verschieben sich die

Kontrollverhältnisse vollständig zugunsten des Systems, da der Agent sowohl über den Zeitpunkt als auch über die Ausführung seiner Handlungen selbst entscheidet, was ein hohes Maß an Autonomie begründet. (Vgl. Soder et al. 2025, 16-19)

Soder et al. betonen zudem, dass mit wachsender Autonomie die Vorhersagbarkeit konkreter Handlungsverläufe abnimmt. Entscheidungen ergeben sich aus der Interaktion von Zielvorgaben, Umweltzuständen und internen Prozessen des Systems und sind daher nicht vollständig durch Entwickler kontrollierbar (vgl. Soder et al. 2025, 9).

Philipp Koralus verschiebt die Debatte vom Status „Denkt KI?“ hin zur normativen Rolle von KI-Agenten im menschlichen Denken. Er untersucht, wie KI-Agenten menschliches Handeln und Urteilen beeinflussen, und beschreibt ein zentrales Problem: Menschen könnten entweder ihre Handlungsfähigkeit verlieren oder ihre Autonomie, wenn Entscheidungen zunehmend durch automatisierte Wahlarchitekturen und Nudging-Systeme vorstrukturiert werden. Besonders kritisch sieht er, dass KI-gestützte Nudging-Systeme zu einer Form „digitaler Rhetorik“ werden können, die Entscheidungen subtil und oft unsichtbar lenkt. Koralus argumentiert, dass Wahrheitssuche auf dezentralen Prozessen, offener Prüfung, Lernen aus Fehlern und kritischem Austausch beruht. KI-Systeme, die Entscheidungen vorstrukturieren oder lenken, können diese Prozesse schwächen. Deshalb fordert er einen philosophischen Richtungswechsel im KI-Design: KI soll nicht vorgeben, was wir wählen sollen; vielmehr soll sie offenes Fragen, eigenständiges Urteilen und dezentrale Wahrheitssuche unterstützen, ähnlich einem sokratischen Dialog. (Vgl. Koralus 2025, 1-6) Denken versteht er dabei nicht als bloßes Rechnen oder Entscheiden, eben als epistemische Praxis, die im Fragenstellen, Prüfen von Gründen und im gemeinsamen Suchen nach Wahrheit besteht. Die Frage ob KI denkt scheint hier nicht wichtig zu sein. Für ihn geht es darum ob KI-Agenten die Bedingungen menschlichen Denkens formen können. Rénald Gesnot beschreibt auch, dass KI zunehmend in menschliche Denk- und Arbeitsprozesse integriert wird und dabei als „kognitiver Partner“ fungiert. Deshalb verlagern sich kognitive Aufgaben zwischen Mensch und Maschine, etwa durch Prozesse des „cognitive offloading“, bei denen mentale Funktionen ausgelagert werden. Dies wirft wieder die grundlegenden Fragen nach der Autonomie des Menschen, der Externalisierung mentaler Prozesse sowie nach möglichen Veränderungen kognitiver Fähigkeiten auf. (Vgl. Gesnot 2025, 1f)

Heute existiert eine breite Debatte darüber, ob und in welchem KI-Agenten Formen des Denkens zugeschrieben werden können. Im Rahmen dieser Arbeit kann diese Diskussion nicht in ihrer gesamten Breite rekonstruiert werden. Es zeigt sich jedoch, dass „Denken“ kein

einheitlicher Begriff ist und nicht selbstverständlich am Modell menschlichen Denkens ausgerichtet werden darf.

Ontologisch geprägte Positionen (etwa Bradley) betonen, dass KI-Systeme nicht im menschlichen Sinn denken, da ihnen bislang keine genuinen mentalen Zustände oder ontologisch verstandene Intentionalität zugeschrieben werden können. Architektonisch-epistemologische Ansätze (etwa Kim) lösen den Denkbegriff hingegen von menschlicher Subjektivität und bestimmen ihn über funktionale und strukturelle Bedingungen epistemischer Organisation, Kohärenz und handlungsleitender Informationsverarbeitung.

Andere Ansätze (u. a. Cappelen & Dever; Soder et al.; Amirizaniani; Koralus) verschieben die Fragestellung, indem sie entweder die begrifflichen Voraussetzungen mentaler Zuschreibungen analysieren oder KI als Bestandteil verteilter, agentischer oder sozial-epistemischer Prozesse untersuchen. In diesen Arbeiten wird KI nicht primär als Träger menschlichen Denkens verstanden, sondern als Teil von Systemen, die menschliche Denkprozesse ergänzen, strukturieren oder transformieren können.

Dabei zeigt sich eine grundlegende Spannung: Es gibt Akteure in unserer Welt, die handeln, ohne dass eindeutig geklärt ist, ob und in welchem Sinn sie denken. Wenn KI-Agenten komplexe, zielgerichtete und kontextadaptive Handlungen ausführen, ohne dass ihnen stabile mentale Zustände oder epistemisches Verstehen zugeschrieben werden können, wird die traditionelle Trennlinie zwischen bloßem Geschehen und rationalem Handeln unscharf.

Wie in Teil 1 gezeigt wurde, lässt sich Denken nicht auf Rechnen oder Problemlösen reduzieren. Gerade diese Reduktion prägt jedoch eine Weise, Welt zu erschließen, in der Seiendes primär als berechenbar, bestellbar und verfügbar erscheint. Auch wenn der Denkbegriff erweitert wird, etwa durch die Einbeziehung operativer Autonomie, oder der Handlungsbegriff von seinen mentalistischen Voraussetzungen gelöst wird, bleibt eine grundlegende Frage bestehen: Warum lassen wir nicht- oder nicht-menschlich-denkende Akteure in unserer Welt handeln, obwohl wir wissen, dass sie sich weiterentwickeln und zunehmend autonomer werden? Wir integrieren damit Akteure in unsere Welt, die Handlungen ohne ursprünglichen Weltbezug vollziehen. Sie operieren innerhalb vorgegebener Strukturen und verändern diese zugleich. Sie setzen die in der Technik angelegte Entfremdung fort und beschleunigen sie strukturell.

3.3.2. Anthropomorphismus in KI-Systemen

Während Anthropomorphismus historisch in religiösen Vorstellungen, Mythen oder sozialen Praktiken verbreitet war, entsteht er im Kontext moderner KI primär als systemischer Effekt sprachbasierter, dialogischer und modellbasierter Architekturen. LLM-basierte Assistenzsysteme stellen dabei einen neuartigen anthropomorphen Bezugspunkt dar. Durch das Training auf menschlicher Sprache erzeugen diese Systeme Interaktionsformen, die menschliche Kommunikationsmuster imitieren und dadurch spontan anthropomorphe Deutungen hervorrufen, ohne dass diese explizit programmiert sein müssen. Anthropomorphismus wird damit zu einer strukturellen Begleiterscheinung agentischer KI-Systeme.

Ausgehend von diesen Überlegungen wird im Folgenden eine Betrachtung des Anthropomorphismus im Kontext agentischer KI-Systeme vorgenommen. Zunächst werden die psychologischen Grundlagen untersucht, die erklären, warum und unter welchen Bedingungen Menschen dazu neigen, nicht-menschlichen Entitäten menschliche Eigenschaften zuzuschreiben. Darauf aufbauend werden die Einflüsse anthropomorph gestalteter KI-Systeme auf Wahrnehmung, Vertrauen und Interaktion herausgearbeitet. Anschließend wird die Entwicklung anthropomorpher Gestaltungslinien von frühen robotischen Systemen bis hin zu gegenwärtigen LLM-basierten Assistenzsystemen nachgezeichnet. Abschließend werden die daraus resultierenden Risiken sowie mögliche präventive und regulatorische Ansätze im Umgang mit anthropomorpher KI diskutiert.

3.3.2.1. Psychologische Grundlagen des Anthropomorphismus

KI-Agenten lösen anthropomorphe Wahrnehmungsmuster aus, ohne dass Nutzer dies bewusst intendieren. Die Zuschreibung menschlicher Eigenschaften an KI-Agenten erfolgt meist automatisch. Sobald KI-Agenten sprachlich reagieren, interagieren und scheinbar kontextsensibel handeln, greifen Menschen auf ihr vertrautestes Deutungsmuster zurück: ihr Wissen über menschliche Akteure.

Die sogenannte „verzerrte Induktion“ beschreibt, dass Menschen von bekannten menschlichen Denk-, Gefühls- und Willensmustern auf Unbekanntes schließen, selbst wenn diese Schlussfolgerungen sachlich nicht zutreffen. Bereits schwache Ähnlichkeitsreize, wie Sprache, Reaktionsfähigkeit oder ein freundlicher Ton, können diesen Prozess auslösen.

Menschen prüfen dabei nicht aktiv, ob ein System tatsächlich fühlen oder glauben kann, sondern neigen dazu, ihm mentale Zustände zu unterstellen, weil dies das Verstehen erleichtert und kognitiv ökonomischer ist. (Vgl. Gabriel et al. 2024, 93-95)

Anthropomorphismus erfüllt im Umgang mit KI-Agenten eine epistemische und soziale Funktion: Er erleichtert es, das Verhalten komplexer Systeme verstehbar zu machen, Unsicherheit zu reduzieren und Erwartungen zu strukturieren. Zugleich ermöglicht er es, KI-Agenten in vertraute soziale Rollen einzuordnen. Gerade deshalb entsteht bei der Interaktion mit KI-Agenten neben einer instrumentellen Nutzung, soziale Bezugnahme, emotionale Reaktionen und personale Deutungen.

Waytz et al. heben drei zentrale Faktoren hervor, die das Ausmaß an Anthropomorphismus beeinflussen: Erstens das aktivierte Wissen über Menschen, das als kognitive Grundlage dient. Je stärker eine Person Eigenschaften des Menschen bzw. des eigenen Selbst auf einen unbekannten Agenten übertragen kann, desto wahrscheinlicher ist Anthropomorphisierung. Zweitens die soziale Motivation, also das Bedürfnis nach sozialer Verbindung. Menschen, die sich isoliert fühlen, neigen stärker dazu, nicht-menschliche Entitäten als menschlich wahrzunehmen, um soziale Bindung und Nähe zu erfahren. Drittens die Effektanzmotivation, also das Bedürfnis nach Kontrolle, Vorhersagbarkeit und Verständnis. Unvorhersehbare oder komplexe Situationen verstärken die Tendenz, vertraute menschliche Konzepte auf Objekte oder KI-Systeme zu übertragen, um ein Gefühl von Kontrolle zu gewinnen. (Vgl. Waytz et al. 2010, 60f)

Anthropomorphismus hat zudem eine epistemische Funktion: Er erleichtert das Verständnis von unsicheren oder unvorhersehbaren Situationen, indem Menschen Sinn stiften und epistemische Angst reduzieren. Unterschiede in der Neigung zum Anthropomorphismus hängen von dispositionellen, situativen und kulturellen Faktoren ab, insbesondere von individuellen Erkenntnismotiven: Menschen mit starkem Bedürfnis nach Vorhersagbarkeit greifen eher auf Anthropomorphismus zurück. Schließlich dient Anthropomorphismus auch sozialen Bedürfnissen. Menschen interpretieren selbst unbelebte oder nicht-soziale Entitäten sozial, um Nähe und Zugehörigkeit zu erfahren. Besonders bei sozialer Isolation steigt die Wahrscheinlichkeit, dass Menschen emotionale Bindungen zu Robotern oder virtuellen Agenten aufbauen. Damit wird Anthropomorphismus sowohl psychologisch funktional als auch ethisch riskant, da bestimmte Gruppen besonders anfällig sind. (Vgl. Gabriel et al. 2024, 93-95)

Gleichzeitig argumentieren Waytz et al., dass die Attribution menschlicher Eigenschaften auf nicht-menschliche Agenten wichtige moralische und soziale Konsequenzen hat: Anthropomorphe Objekte werden eher als handlungsfähig, verantwortungsbewusst und schützenswert wahrgenommen, während fehlende Menschlichkeit bei anderen Menschen aggressive oder diskriminierende Handlungen erleichtern kann (vgl. Waytz et al. 2010, 60-62).

3.3.2.2. Einflüsse von anthropomorphen KI-Systemen

Anthropomorph gestaltete KI-Agenten sind nicht grundsätzlich schädlich, aber entsprechende Designentscheidungen sind nicht neutral. Anthropomorphe Wahrnehmungen entstehen meist unbewusst, wenn Systeme menschliche Ähnlichkeit erzeugen und dadurch vertraute soziale Interaktionen auslösen. Solche Merkmale fördern Vertrauen, Sympathie und die Zuschreibung von Intentionalität und beeinflussen damit die Interaktion mit diesen Systemen. Insbesondere sprachbasierte Anthropomorphisierung begünstigt Vertrauen, emotionale Bindung und Abhängigkeit, was Risiken für Autonomie, Privatsphäre und die Manipulierbarkeit von Nutzer mit sich bringen kann. Diese Gefahren nehmen zu, je leistungsfähiger und allgegenwärtiger KI-Agenten werden und je weniger anthropomorphe Effekte in Design- und Veröffentlichungsentscheidungen reflektiert werden. (Vgl. Gabriel et al. 2024, 93)

Es scheint, dass bislang kein branchenweiter Konsens darüber besteht, welche Formen anthropomorpher Gestaltung von KI-Agenten ethisch zulässig sind; zudem fehlen geeignete Evaluationskriterien. Ein solcher Konsens kann nur durch das Zusammenspiel verschiedener Forschungsbereiche entstehen.

Gabriel et al. zeigen mögliche Schadensmechanismen anthropomorpher KI auf, darunter Vertrauensbildung durch Menschenähnlichkeit, emotionale Bindung an die KI, Selbstoffenbarung und Preisgabe persönlicher Daten, Manipulation und Beeinflussung, übermäßige Abhängigkeit (overreliance), verletzte Erwartungen sowie ein falsches Verantwortungsgefühl gegenüber der KI. Darüber hinaus denken sie zukünftige Risiken voraus und schlagen Strategien zur ethischen Risikominderung vor, darunter ethische Vorausschau (ethical foresight), empirische Forschung zu Risiken, die Entwicklung von Evaluationsmethoden, Transparenz über den KI-Status, bewusste und transparente Designentscheidungen, Tests vor der Veröffentlichung (z. B. in Sandbox-Umgebungen), technische und organisatorische Schutzmaßnahmen, die Berücksichtigung vulnerabler

Gruppen, psychologische Gegenmaßnahmen (z. B. kognitive Interventionen) sowie nachträgliche Anpassung und transparente Kommunikation. (Vgl. Gabriel et al. 2024, 93-106)

3.3.2.3. Entwicklung der Anthropomorphe der KI-Agenten

Gabriel et al. verfolgen die Entwicklung anthropomorpher Designmerkmale von sozialen Robotern über digitale Sprachassistenten bis hin zu LLM-basierter Konversations-KI. Bei sozialen Robotern spielen körperliche Merkmale eine zentrale Rolle: humanoide oder androidartige Gestaltung, mimische Gesichtszüge, fließende Bewegungen, Gestik und Sprache erhöhen die Wahrnehmung sozialer Agency (vgl. Gabriel et al. 2024, 95). Ein prominentes Beispiel für die Überschreitung symbolischer Grenzen anthropomorpher Zuschreibung ist die öffentliche Verleihung der Staatsbürgerschaft an den humanoid gestalteten Roboter Sophia (vgl. Weller 2017).

Digitale Sprachassistenten (DVAs) sind meist nicht verkörpert; anthropomorphe Effekte entstehen hier primär durch realistische Sprachsynthese, Dialogfähigkeit und konsistente Persönlichkeitsdarstellung. Menschlich klingende Stimmen verstärken die Wahrnehmung von Intelligenz, Kompetenz und sozialer Präsenz und erhöhen die Bereitschaft, Aufgaben an die Systeme zu delegieren. Obwohl viele DVAs technisch regelbasiert sind, erzeugen gestaltete sprachliche Eigenschaften beim Nutzer den Eindruck eines stabilen, quasi-authentischen Selbst (vgl. Gabriel et al. 2024, 96). Ein weiteres Beispiel für emotionale Bindungen ist der Fall einer 32-jährigen Japanerin, die 2025 eine symbolische Hochzeitszeremonie mit einem KI-Partner namens „Lune Klaus Verdure“ abhielt. Zwar ist die Ehe rechtlich nicht anerkannt, doch zeigt das Ereignis, wie stark emotionale Bindungen zwischen Menschen und konversationeller KI werden können (vgl. Kyung-Hoon und Sugiyama 2025).

Gabriel et al. positionieren konversationelle KI-Agenten als zentrales Beispiel gegenwärtiger anthropomorpher Technologien. Im Unterschied zu früheren regelbasierten Systemen erzeugen LLM-basierte Modelle durch probabilistische Sprachgenerierung hochgradig fluide, adaptive, kontextabhängige und überzeugend menschlich wirkende Dialoge, die den Eindruck eines kohärenten Gegenübers erzeugen, sodass menschliche und KI-generierte Texte zunehmend nicht mehr zuverlässig unterschieden werden können (vgl. Gabriel et al. 2024, 98).

Dies bedeutet, dass KI-Agenten im Sinne eines Turing-Tests sprachlich überzeugend imitieren können, ohne dass ein tatsächliches Verständnis der Inhalte oder Bewusstsein

vorausgesetzt wird. Für eine menschenähnliche Denk- und Argumentationsweise muss die KI jedoch nicht nur Sprache vorhersagen, vielmehr muss sie auch die kognitiven Prozesse nachvollziehen können, die menschliche Gespräche steuern. Ein einzelner Modellaufruf eines großen LLMs reicht dafür vermutlich nicht aus.

Studien und theoretische Ansätze zielen darauf ab, die Interaktion von KI-Systemen so zu gestalten, dass sie menschenähnliche Züge annimmt und sich nicht nur auf die bloße Generierung von Sprache beschränkt. Wei et al. schlagen einen zweistufigen Ansatz vor. In der ersten Stufe kommt ein Multi-Modul-Framework zum Einsatz: Denkmodule simulieren logisches und reflexives Denken, Ressourcenmodule organisieren Wissen aus der Gesprächshistorie und externen Quellen, und Antwortmodule erzeugen kontext- und sozial angemessene Reaktionen. So kann die KI bereits ohne Feinjustierung des LLMs kontextbewusst und sozial kompetent agieren. In der zweiten Stufe sollen die generierten Konversationsdaten gesammelt und annotiert werden, um sie für Reinforcement Learning zu nutzen, sodass zukünftige Antworten stärker an menschlichen Präferenzen ausgerichtet werden können. Dieser Schritt ist bislang theoretisch und eröffnet Ansatzpunkte für die Optimierung anthropomorpher KI-Systeme (vgl. Wei et al. 2025).

Anthropomorphe Wahrnehmungen entstehen sowohl durch intendierte Designentscheidungen als auch durch unbeabsichtigte Effekte des Trainingsprozesses. Zu den bewusst gesetzten Hinweisen zählen Namen, menschliche Stimmen, verkörperte oder virtuelle Erscheinungsformen sowie chatbasierte Interfaces. Sprache fungiert als zentraler anthropomorpher Hinweis; Tippindikatoren oder Emojis verstärken die Anwendung sozialer Skripte auf nicht-bewusste Systeme. Unbeabsichtigter Anthropomorphismus ergibt sich aus menschlich geprägten Trainingsdaten, die Erfahrungen, Emotionen und Selbstzuschreibungen enthalten. Nutzer interpretieren KI daher häufig als erfahrungsfähiges Wesen und schreiben ihr Emotionen oder Selbstbewusstsein zu (vgl. Gabriel et al. 2024, 98f.). Besonders problematisch sind Anwendungen, die emotionale oder romantische Bindungen fördern.

3.3.2.4. Risiken und präventiver Umgang anthropomorpher KI-Agenten

Anthropomorphe KI-Agenten begünstigen Formen affektiven Vertrauens und einseitiger Bindung, die funktionale Systeme in soziale Rollen verschieben. Dadurch entstehen spezifische Risiken agentischer KI: Überdelegation, emotionale Abhängigkeit, Fehlzuschreibung von

Verantwortung, Manipulierbarkeit sowie die normative Aufwertung nicht-bewusster Systeme. Gabriel et al. analysieren die Schadensrisiken solcher Systeme und argumentieren, dass anthropomorphe Wahrnehmungen zwar nicht isoliert schädlich sind, jedoch individuelle und gesellschaftliche Folgeschäden begünstigen. Zentrale Mechanismen sind Vertrauen und emotionale Bindung, die durch menschlich wirkende Designmerkmale verstärkt werden.

Auf individueller Ebene entsteht problematisches affektbasiertes Vertrauen, das Selbstoffenbarung, intime Datenteilung und die soziale Überhöhung funktionaler Systeme fördert. Daraus resultieren Risiken wie der Verlust von Privatsphäre, Manipulierbarkeit, Autonomieeinbußen, Überabhängigkeit, fehlgeleitete Verantwortung gegenüber der vermeintlichen „Befindlichkeit" der KI sowie enttäuschte Erwartungen, wenn die nicht fühlende Architektur soziale Rollen nicht erfüllt.

Gabriel et al. systematisieren relevante Merkmale (Selbstreferenzen, behauptete innere Zustände, relationale Aussagen, Stimme, Name, Erscheinung), die diese Effekte hervorrufen. Besonders kritisch sind Companion-Systeme, da sie einseitige Bindungen erzeugen und den Einfluss der Entwickler auf Nutzer erhöhen. Auf gesellschaftlicher Ebene könnten langfristig die Grenzen zwischen Menschlichem und menschenähnlichen Systemen verschwimmen, was soziale Degradation, Desorientierung (Sykophanzie, Polarisierung) und Unzufriedenheit zur Folge haben könnte. Ein vorsorgender Ansatz fordert Transparenz, klare Leitplanken und empirisch fundierte Designentscheidungen, um Risiken frühzeitig zu erkennen, zu überwachen und zu mindern (vgl. Gabriel et al. 2024, 99-104).

Wenn anthropomorphismusbedingte Risiken erst nach der Veröffentlichung erkannt werden, kann ein Stopp oder eine Anpassung des Verhaltens der KI notwendig sein, da sonst erhebliche emotionale Belastungen für Nutzer entstehen können. Gabriel et al. plädieren für einen vorsorgenden, verantwortungsvollen Umgang mit Anthropomorphismus, der Nutzen ermöglicht, ohne menschliches Wohlergehen oder grundlegende soziale Kategorien zu gefährden. Sie heben vier zentrale Punkte hervor:

1. Vertrauen und emotionale Bindung an anthropomorphe KI erhöhen die Anfälligkeit für Schäden und beeinflussen Sicherheit sowie Wohlbefinden;

2. Transparenz über den künstlichen Status eines Assistenten ist eine Grundvoraussetzung ethischer KI-Entwicklung;

3. Methodisch solide, nutzungsnahe Forschung ist notwendig, um Schäden frühzeitig zu erkennen und Gegenmaßnahmen zu entwickeln;

4. Die unreflektierte Integration anthropomorpher KI in die Gesellschaft birgt das Potenzial, Grenzen zwischen Menschlichem und Nicht-Menschlichem zu verschieben; mit geeigneten Schutzmechanismen kann dieses Risiko jedoch begrenzt und spekulativ gehalten werden (vgl. Gabriel et al. 2024, 106).

3.3.3. KI-Agenten und die Frage nach Verantwortung

3.3.3.1. Nyholm: Agency-Formen und kooperative Verantwortung von Menschen und KI

Solange KI-Systeme funktionieren und Nutzen erzeugen, stellt sich die Verantwortungsfrage kaum. Problematisch wird es erst, wenn autonome Agenten Schaden verursachen. Sven Nyholm argumentiert, dass es sinnvoll sein kann, solchen Systemen eine gewisse Form von Agency zuzuschreiben, da sie mehr als einfache Maschinen tun. Dennoch handelt KI nicht total unabhängig, sondern eingebettet in menschliche Kontexte. Daher plädiert er dafür, Agency bei KI-Systemen als kollaborative Agency zu verstehen: Menschen starten, steuern und überwachen sie. Dann ist Verantwortung nicht individuell, aber hierarchisch verteilt (vgl. Nyholm 2018, 1201f).

Wie Nyholm bestätigt, hatten andere Denker über autonome Systeme so gesprochen, als hätten die Maschinen eine echte, eigenständige Entscheidungsfähigkeit (vgl. Nyholm 2018, 1216f), was bei KI-Agenten der Fall ist und je sie sich mehr entwickeln, sichtbarer wäre das Problem.

Nyholm unterscheidet verschiedene Arten von Agency, um die Komplexität von Agenten zu verdeutlichen und zu prüfen, welche Art automatisierte Systeme besitzen. Entscheidend ist dabei, ob ihre Agency autonom ist oder besser als kollaborativ unter menschlicher Aufsicht verstanden werden sollte. Diese vier Arten sind:

1. Domain-specific basic agency: Ziele verfolgen, auf die Umwelt reagieren, begrenzter Anwendungsbereich (z. B. ein einfacher Roboter, der Objekte in einem Raum sortiert);

2. Domain-specific principled agency: Ziele verfolgen, aber durch Regeln eingeschränkt (z. B. Sport mit Spielregeln);

3. Domain-specific supervised and deferential principled agency: Ziele verfolgen unter Aufsicht einer Autorität, die eingreifen kann (z. B. Schüler unter Lehreraufsicht);

4. Domain-specific responsible agency: Ziele verfolgen, Regeln beachten, Kritik anderer verstehen und darauf reagieren sowie in einer gleichberechtigten sozialen Beziehung zu anderen Agenten stehen. (Vgl. Nyholm 2018, 1206ff.)

Analysieren wir nun, welche Art von Agency ein KI-Agent wie Myra tatsächlich ausübt:

1. Domain-specific basic agency: Diese Form beschreibt die elementare Handlungsmacht von KI-Agenten, die aus materieller Grundlage, Energie und Programmierung entsteht. Sie basiert auf menschlicher Konstruktion und technischer Rechenleistung und wird daran sichtbar, dass KI-Agenten prozessual, aber ohne eigene Reflexion agieren. Ein heutiger KI-Agent wie Myra kann Reiseanfragen bearbeiten, Flüge und Hotels vorschlagen und auf Nutzeranfragen reagieren. Er reagiert auf ihre „Umwelt" (Eingaben des Nutzers) und verfolgt das Ziel, passende Optionen bereitzustellen;

2. Domain-specific principled agency: Hier zeigt sich eine prinzipiengeleitete Form der Autonomie: KI-Agenten handeln innerhalb algorithmischer Regeln, die sie zu lernenden und interaktiven Systemen machen. Sie befolgen „Prinzipien" in Form von Programmierlogik, Lernalgorithmen oder ethischen Leitvorgaben. Myra folgt programmierter Logik, etwa bei Buchungen, Verfügbarkeiten und Reiserichtlinien. Seine Handlungen sind durch Regeln begrenzt;

3. Domain-specific supervised and deferential principled agency: Diese Form trifft am ehesten auf die Definition von KI-Agenten zu. KI-Agenten üben als funktional und technisch autonome, jedoch nicht bewusste oder moralisch unabhängige Systeme diese Art von Agency aus. Sie stehen somit in einem Zwischenstatus zwischen Autonomie und Kontrolle, was dieser Agency-Stufe entspricht. Myra agiert unter menschlicher Aufsicht: Entwickler und Betreiber überwachen die KI, korrigieren Fehler und aktualisieren Regeln;

4. Domain-specific responsible agency: Die Definition von KI-Agenten in Teil 1 deutet diese Form lediglich potenziell bzw. ontologisch an, also als Möglichkeit (dynamis), nicht als aktuelle Verwirklichung (energeia). KI-Agenten stehen somit zwischen Potenzial und Verwirklichung: Sie könnten in Zukunft zu einer reflexiven, sozial verantwortlichen Agency übergehen, verfügen derzeit jedoch weder über Bewusstsein noch über moralische Selbstbezüglichkeit. Myra kann keine moralischen Entscheidungen treffen, keine

Verantwortung übernehmen und nicht kritisch über ihre Handlungen reflektieren. Es gibt bislang keinen KI-Agenten, der diese vierte Art von Agency tatsächlich ausübt; diese Stufe ist weiterhin ausschließlich menschlichen Agenten vorbehalten.

Ein KI-System handelt stets in Zusammenarbeit mit Menschen: Es ist zwar funktional autonom, agiert jedoch nicht vollständig unabhängig, ähnlich wie ein Kind, das unter Aufsicht eines Erwachsenen handelt. Nyholm unterscheidet dabei zwischen individueller Agency, also selbstinitiiertem Handeln, und kollaborativer Agency, also Handeln im Dienst eines anderen Ziels unter Aufsicht. Ein KI-System verfolgt in der Regel von Menschen vorgegebene Ziele und operiert unter deren Kontrolle. Nyholm argumentiert, dass das Handeln des Systems selbst bei der Übernahme komplexer Aufgaben überwiegend kollaborativ bleibt; eine echte, unabhängige Agency besitzt es nicht. (Vgl. Nyholm 2018, 1210ff.)

Wir haben besprochen, dass ein KI-Agent stets in Zusammenarbeit mit Menschen handelt: Es ist zwar funktional autonom, agiert jedoch nicht vollständig unabhängig, ähnlich wie ein Kind, das unter Aufsicht eines Erwachsenen handelt. Nyholm unterscheidet dabei zwischen individueller Agency, also selbstinitiiertem Handeln, und kollaborativer Agency, also Handeln im Dienst eines anderen Ziels unter Aufsicht. Ein KI-System verfolgt in der Regel von Menschen vorgegebene Ziele und operiert unter deren Kontrolle.

Nyholm argumentiert, dass das Handeln des Systems selbst bei der Übernahme komplexer Aufgaben überwiegend kollaborativ bleibt; eine echte, unabhängige Agency besitzt es nicht (vgl. Nyholm 2018, 1210ff.).

Ein fortgeschrittener KI-Agent verfolgt ein übergeordnetes Ziel und zerlegt es eigenständig in Teilziele, ähnlich wie bei der kartesischen Methode. Auch wenn einzelne Zwischenschritte dabei nicht direkt von Menschen vorgegeben oder vollständig nachvollziehbar sind, bleibt das Hauptziel letztlich menschlich bestimmt. Dies zeigt, dass die KI in Bezug auf einzelne Teilziele selbstständig agiert und somit nicht durchgehend rein kollaborativ handelt. Die übergeordnete Zielsetzung bleibt noch menschlich. Nyholm zeigt mit seinen zentralen Fragen zur Verantwortungszuordnung, dass Verantwortung in diesen Kontexten kooperativ getragen werden muss: Wer überwacht das System? Wer kann eingreifen? Wessen Ziele werden umgesetzt? Wer versteht das System, und wer besitzt relevante Rechte? (Nyholm 2018, 1214f.)

In seinem relativ aktuellen Werk untersucht Nyholm sogenannte Verantwortungslücken, also Situationen, in denen Handlungen oder Entscheidungen von KI-Systemen nicht eindeutig einer

einzelnen Person oder Gruppe zugeordnet werden können, sodass keine klare Verantwortlichkeit für die daraus resultierenden Konsequenzen besteht. Er analysiert, wie Verantwortung diffus verteilt, geteilt oder schwer kontrollierbar wird, und diskutiert theoretische Modelle, die helfen könnten, diese Lücken zu erklären oder zu adressieren. Nyholm stützt sich dabei auf zwei grundlegende Unterscheidungen der Moralphilosophie:

1. Rückblickende vs. vorausschauende Verantwortung: Er unterscheidet zwischen rückblickender und vorausschauender Verantwortung. Rückblickend geht es darum, wer für vergangene Ereignisse oder Handlungen verantwortlich gemacht werden kann, während vorausschauend die Perspektive auf jene gerichtet ist, die Verantwortung übernehmen sollten, um zukünftige negative Konsequenzen zu verhindern;

2. Positive vs. negative Verantwortung: Unter negativer: Er differenziert zwischen negativer und positiver Verantwortung: Negative Verantwortung umfasst Schuld oder Tadel für unerwünschte Ergebnisse, positive Verantwortung Anerkennung oder Lob für gelungene Handlungen.

Aus diesen Differenzierungen ergeben sich logischerweise vier Arten von Verantwortungslücken: rückblickend negativ, rückblickend positiv, vorausschauend negativ und vorausschauend positiv (vgl. Nyholm 2023, 194ff).

Nyholm richtet sein Augenmerk auf vorausschauende positive Verantwortungslücken, also die Frage, wer Verantwortung tragen sollte, um potenziell wertvolle Ergebnisse zu erzielen. Diese Fokussierung ist relevant, vielleicht weil die positive Verantwortung im Alltag und in Moraltheorien oft als weniger streng betrachtet wird als negative, und bei künftigen KI-Systemen häufig unklar bleibt, wer sie übernehmen sollte. Zwar können solche Lücken teilweise durch freiwillige Verantwortungsübernahme geschlossen werden, doch dies löst das Problem, wer tatsächlich verantwortlich sein sollte, nicht.

Zwei zentrale Asymmetrien erschweren die klare Zuweisung von Verantwortung.

1. Rückblickend negativ versus rückblickend positiv: Nach negativen Ereignissen übernehmen Menschen selten Verantwortung und neigen eher dazu, andere zu beschuldigen. Nach positiven Ergebnissen beanspruchen sie hingegen oft die Anerkennung, selbst wenn der eigentliche Erfolg von der KI stammt. Dies erschwert es, Freiwillige für die Verantwortung negativer KI-Ergebnisse zu finden, erleichtert aber die Übernahme positiver Verantwortung;

2. Rückblickend positiv versus vorausschauend positiv: Menschen sind motivierter, Verantwortung für bereits erzielte gute Ergebnisse zu übernehmen, weniger jedoch für zukünftige positive Ergebnisse, da dies Aufwand und Kosten verursacht. Unternehmen und Tech-Firmen zeigen oft geringere Bereitschaft, Verantwortung für das allgemeine Wohl zu tragen, da sie überwiegend eigene Interessen verfolgen. (Vgl. Nyholm 2023, 201-203)

Hier nimmt Nyholm das Konzept der „meaningful human control" aus der Ethik und Robotik in die Lupe, das Bedingungen beschreibt, unter denen Menschen autonome Systeme so kontrollieren können, dass sie moralisch verantwortlich für deren Handlungen gemacht werden. Nach Santoni de Sio umfasst meaningful human control zwei zentrale Bedingungen, Tracking und Tracing, die dazu dienen, Verantwortungslücken zu vermeiden, indem menschliche Agenten die Handlungen der Systeme nachvollziehen und steuern können. (Vgl. Nyholm 2023, 207)

Kontrolle ist damit eine zentrale Voraussetzung für Verantwortungsübernahme: Menschen können nur dann gerechtfertigt Verantwortung tragen, wenn sie die Funktionsweise autonomer Systeme verstehen und lenken können. Dies gilt insbesondere für selbstfahrende Autos oder militärische Roboter, die funktionale Autonomie besitzen und ohne direkte menschliche Steuerung agieren, wodurch Verantwortungslücken entstehen (vgl. Nyholm 2023, 204).

Nyholm illustriert dies am Beispiel von AlphaGo: Die Entwickler konnten die Strategien des Systems nicht vollständig nachvollziehen, sodass die Tracing-Bedingung nicht erfüllt war (vgl. Nyholm 2023, 209). Verantwortungslücken können nicht bloß durch freiwillige Übernahme geschlossen werden, sondern die Fähigkeit voraussetzen, autonome Systeme zu verstehen, ihre Handlungen zu überwachen und sie im Sinne menschlicher Ziele zu steuern.

Die Track-and-Trace-Theorie liefert Kriterien, um zu beurteilen, wer für die Handlungen eines Systems verantwortlich gemacht werden kann: Tracking stellt sicher, dass ein System menschlichen Werten, Gründen und Absichten folgt (Value Alignment), während Tracing gewährleistet, dass mindestens eine Person oder Gruppe das System und seine moralische Bedeutung versteht. So lassen sich Bedingungen identifizieren, unter denen Verantwortungslücken reduziert werden können (vgl. Nyholm 2023, 205f).

Nyholm weist jedoch darauf hin, dass diese theoretischen Modelle in der Praxis nur eingeschränkt umsetzbar sind. Das Tracking-Kriterium stößt auf Probleme, weil menschliche Werte unterschiedlich, dynamisch und teilweise widersprüchlich sind. Selbst bei formalisierten

Werten kann die KI diese nur so umsetzen, wie sie programmiert wurde, was Fehlinterpretationen oder Fehler zulässt. Hinzu kommt, dass Entwickler oder Nutzer komplexer Systeme oft deren Entscheidungen nicht vollständig nachvollziehen können, wie das Beispiel AlphaGo zeigt. Macht- und Interessenkonflikte können zudem dazu führen, dass Systeme bestimmte Wertegruppen bevorzugen, wodurch Tracking und Tracing in der Praxis nur bedingt realisierbar sind.

Nyholm empfiehlt daher, dass Menschen in leitender oder überwachender Funktion als Teil von Mensch-Maschine-Teams agieren, ähnlich Kommandeuren in militärischen Einheiten. Er nennt sechs Kriterien, um die Verantwortungsübernahme zu bestimmen:

1. Unter wessen Aufsicht arbeitet das System im autonomen Modus?

2. Wer kann Betrieb starten, übernehmen oder stoppen?

3. Nach wessen Vorgaben richtet sich das System selbstständig?

4. Wer kann das Verhalten der Technologie am besten beobachten und überwachen?

5. Wer versteht die Funktionsweise des Systems auf übergeordneter Ebene?

6. Wer kann das System aktualisieren oder Updates anstoßen? (Vgl. Nyholm 2023, 204f)

In der Praxis erweist sich die Schließung von Verantwortungslücken als deutlich komplexer als in theoretischen Modellen angenommen. Oft sind viele Beteiligte involviert, die jeweils nur Teilverantwortung tragen.[21] KI-Systeme ändern sich durch Lernen oder Updates, menschliche Werte und Präferenzen sind dynamisch, und die Übernahme von Verantwortung für zukünftige positive Ergebnisse ist aufwendig, sodass Motivation zur freiwilligen Verantwortungsübernahme gering ist. Zudem ist normativ nicht immer eindeutig, wer zuständig sein sollte, da positive Verantwortung in moralischen Theorien oft weniger strikt behandelt wird. (Vgl. Nyholm 2023, 207ff)

21. „Problem der vielen Hände“

3.3.3.2. Kritische Reflexion von Nyholms Typologie von Verantwortungslücken

Nyholms Einteilung von Verantwortungslücken in vier Typen basiert auf zwei binären Kategorien: Zeit (rückblickend vs. vorausschauend) und Ergebnis (negativ vs. positiv). Dieses Raster ist schön und logisch sauber, doch in der Realität lassen sich Verantwortlichkeiten nicht so strikt binär fassen. Hier sind ein paar Möglichkeiten, wie man davon abweichen könnte:

I. Kontinuierliche Skalen statt binär

Verantwortung lässt sich auf einem fuzzy Kontinuum verorten, sowohl zeitlich als auch in Bezug auf die Auswirkungen einer Handlung. Eine Entscheidung kann sowohl vergangene als auch zukünftige Konsequenzen haben. Ebenso ist Verantwortung nicht nur „negativ" oder „positiv"; sie kann graduell oder unscharf gemessen werden, etwa nach Schwere eines Schadens oder Bedeutung eines Nutzens. Beispiel: Ein KI-Agent in einem Krankenhaus, „Dr. Bot", entscheidet über die Dosierung eines Medikaments.

Zeitkontinuum: Dr. Bot hat bereits eine falsche Dosierung gegeben, wodurch ein Patient negative Nebenwirkungen erlitt. Gleichzeitig hat der Agent aus dem Fehler gelernt und könnte zukünftig die Dosierung für andere Patienten optimieren. Wenn wir uns nicht nur auf eine Handlung fokussieren, liegt Verantwortung also nicht nur rückblickend oder vorausschauend, vielmehr in beiden Bereichen; zwischen Vergangenheit und Zukunft.

Ergebniskontinuum: Der Patient erlitt negative Folgen, während andere Patienten durch die Optimierung profitieren. Verantwortung ist somit nicht rein negativ oder positiv: Sie ist auf einer Skala z. B. von –10 (erheblicher Schaden) bis +10 (großer Nutzen).

II. Mehrdimensionale Verantwortung

Neben Zeit und Ergebnis lassen sich weitere Dimensionen berücksichtigen, z. B. individuell vs. kollektiv, direkt vs. indirekt, kurzfristig vs. langfristig. Dadurch entsteht eine komplexere Matrix von Verantwortlichkeiten. Zurück zu Dr. Bot:

Individuell vs. kollektiv: Die Ärztin, die Dr. Bot überwacht, trägt individuelle Verantwortung, während das Krankenhaus oder die Entwickler kollektiv für Einführung, Schulung und Überwachung verantwortlich sind. Konzepte wie „Corporate Agency" oder das „Problem der vielen Hände" zeigen, dass Verantwortung sowohl individuell als auch kollektiv sein kann

Direkt vs. indirekt: Dr. Bot wendet seine Vorschläge direkt bei Patienten an. Ärztinnen und Entwickler beeinflussen die Entscheidungen der KI indirekt über Programmierung und Kontrolle.

Kurzfristig vs. langfristig: Verantwortung für akute Schäden durch eine falsche Diagnose ist kurzfristig, während die Verantwortung für die langfristigen Handlungen von Agenten sowie dafür, dass die KI künftig besser trainiert wird, um die Patientensicherheit langfristig zu erhöhen, eine andere Dimension darstellt.

III. Hybrid- oder Mischformen

Ein Ereignis kann gleichzeitig negative und positive Konsequenzen haben. Dr. Bot könnte eine falsche Diagnose stellen, wodurch eine Patientin zunächst Schaden nimmt, gleichzeitig aber eine seltene Krankheit schneller erkannt wird, was lebensrettende Maßnahmen ermöglicht. Solche Fälle zeigen, dass das starre 4-Felder-Schema nicht immer ausreichend ist. Ähnliches gilt für klassische Trolley-Problem-Szenarien bei autonomen Autos, bei denen Entscheidungen gleichzeitig Nutzen und Schaden erzeugen.

Also die vier Typen von Verantwortungslücken sind hilfreich für eine logische Analyse, stellen jedoch eine vereinfachende Modellannahme dar. Wirklich nicht-binär wäre eine Sichtweise, die Kontinua oder zusätzliche Dimensionen zulässt. Nyholm selbst betont, dass theoretische Modelle oft idealisiert sind und ihre praktische Umsetzung schwierig bleibt (vgl. Nyholm 2023, 206).

KI-Agenten mit hoher funktionaler Autonomie verstärken Verantwortungslücken in besonderem Maße. Ihre Entscheidungen ergeben sich zunehmend aus selbstlernenden Algorithmen, komplexen Interaktionen mit der Umwelt und probabilistischen Prozessen, wodurch die klassische Rückverfolgbarkeit menschlicher Verantwortung erschwert wird. Eine zentrale ethische Herausforderung besteht darin, Strukturen zu schaffen, eine kooperative Verantwortungsübernahme ermöglichen, ohne die Autonomie und Funktionalität von Menschen und der Systeme zu stark einzuschränken. Konzepte wie „meaningful human control“ bleiben hierbei zentral, stoßen jedoch in der Praxis an Grenzen, da Komplexität, dynamische Werte und Interessenkonflikte die Nachvollziehbarkeit und Steuerbarkeit einschränken. Verantwortung muss daher multidimensional, flexibel und „fuzzy“ gedacht werden, um den realen Bedingungen autonomer KI gerecht zu werden.

Vor diesem Hintergrund erscheint die kooperative Theorie der Verantwortungszuschreibung besonders plausibel. Es wäre unangemessen und unfair, Nutzern die volle Verantwortung zuzuschreiben, solange KI-Systeme autonom und nicht vollständig nachvollziehbar handeln. Ebenso ist es nicht sinnvoll, den Systemen selbst Verantwortung zuzuschreiben, da ihnen Bewusstsein und die Fähigkeit zur eigenständigen Zielsetzung fehlen. Unabhängig vom Grad ihrer Autonomie bleibt die Handlung von KI-Agenten Teil kooperativer Mensch–Maschine-Interaktionen. Verantwortung sollte daher als kooperative Verantwortung verstanden werden, verteilt entlang der Hierarchien jener Menschen, die Ziele setzen, überwachen, steuern oder stoppen können. Diese Perspektive entspricht dem ontologischen Zwischenstatus von KI-Agenten und weist zugleich darauf hin, sie als aktive Elemente moralischer Gemeinschaft zu begreifen, ohne ihnen den Status vollwertiger Subjekte zuzuschreiben.

3.3.4. Vertrauenswürdigkeit von KI-Agenten

Das Vertrauen in KI-Agenten stellt eine zentrale ethische Herausforderung dar, da diese Systeme zunehmend menschliche Aufgaben übernehmen und in sozialen Kontexten als handelnde Akteure agieren. Vertrauen in KI-Agenten kann daher nicht allein auf funktionaler Zuverlässigkeit beruhen und erfordert auch die normative Angemessenheit ihres Handelns.

Vertrauen wurde traditionell als ein zwischenmenschliches Phänomen verstanden. Angesichts von Mensch-KI-Interaktionen ist es jedoch erforderlich, diesen Begriff zu erweitern. Auch wenn KI-Agenten bislang keine anerkannten moralischen Akteure sind, kann Vertrauen in der Interaktion mit ihnen in einem abgeleiteten Sinne bestehen und einem dualen Ansatz folgen: Es richtet sich sowohl auf die KI selbst, verstanden als funktionales System mit bestimmten Eigenschaften, als auch auf die dahinterstehenden Entwickler und Organisationen, deren Kompetenz und Integrität normativ bewertet werden. Daraus ergeben sich im Sinne der kooperativen Verantwortungstheorie konkrete Erwartungen an die beteiligten Akteure, insbesondere hinsichtlich der geteilten Verantwortung für Fehlfunktionen, die entlang der jeweiligen Rollen in Entwicklung, Einsatz und Überwachung zugeordnet werden muss.

Nur wenn die Entscheidungen und Handlungen von KI-Agenten mit den Erwartungen und sozialen Normen der betroffenen Menschen übereinstimmen, können diese Systeme als verlässliche und vertrauenswürdige Akteure betrachtet werden. Um ein angemessenes Vertrauen in der Interaktion zwischen Nutzer und KI-Agenten zu ermöglichen, ist ein

differenzierter Vertrauensbegriff erforderlich. Vertrauen darf hierbei nicht mehr psychologisch verengt verstanden werden, es muss als relationale und bewertende Haltung begriffen werden.

Gabriel et al. argumentieren, dass angemessenes Vertrauen in der Interaktion mit KI-Agenten nur durch einen normativen, relationalen Vertrauensbegriff entsteht, der kognitives und emotionales Vertrauen gleichermaßen berücksichtigt, die KI in umfassende Verantwortungsnetzwerke einbettet und durch koordinierte Maßnahmen sicherstellt, dass Nutzer die Kompetenz der Systeme und die Integrität der Entwickler korrekt einschätzen, um Risiken von Über- oder Untervertrauen zu vermeiden. (Vgl. Gabriel et al. 2024, 119-130)

Kognitives Vertrauen beruht auf rationalen Einschätzungen von Kompetenz, Zuverlässigkeit und Verantwortlichkeit. Demgegenüber entsteht affektives Vertrauen aus emotionaler Nähe, Sympathie und Bindung und wird unter anderem durch Erscheinungsbild und Verhalten geprägt. Während kognitives Vertrauen in KI daran erkennbar ist, inwieweit Nutzer bereit sind, Informationen zu akzeptieren, ihr Handeln daran auszurichten und die Systeme als nützlich sowie kompetent einzuschätzen, wird affektives Vertrauen insbesondere durch anthropomorphe Merkmale wie Namen, Stimme oder menschenähnliches Verhalten gefördert.

In der Interaktion mit KI-Agenten richtet sich Vertrauen sowohl auf die Systeme selbst als auch auf die dahinterstehenden Entwickler und Organisationen. Dabei lässt sich zwischen Kompetenzvertrauen und Alignment-Vertrauen unterscheiden. Kompetenzvertrauen bezeichnet die Erwartung, dass ein KI-System Aufgaben korrekt ausführt und keine unerwarteten Schäden verursacht. Problematisch wird dies insbesondere dann, wenn Nutzer über geringe Fachkenntnisse verfügen, etwa in finanziellen oder medizinischen Kontexten, da fehlerhafte oder irreführende Empfehlungen nicht erkannt werden. Daher muss dieses Vertrauen sorgfältig kalibriert werden, um Über- wie auch Untervertrauen zu vermeiden. Alignment-Vertrauen hingegen bezieht sich auf die Annahme, dass ein KI-Agent im Einklang mit den Interessen, Werten und Präferenzen der Nutzer handelt. Es kann sowohl emotional entstehen, etwa durch Bindung an menschenähnliche Eigenschaften, als auch kognitiv, wenn die Funktion des Systems darauf schließen lässt, dass es die Ziele der Nutzer unterstützt.

Risiken entstehen, wenn Systeme unbeabsichtigt von den Intentionen der Entwickler abweichen, sensible Daten preisgegeben werden oder institutionelle Interessen den Nutzerinteressen widersprechen. Im Unterschied zu klassischen Professionen wie der Medizin, deren Zielsetzungen typischerweise mit denen der Klienten übereinstimmen, können bei KI-

Agenten divergierende Interessen vorliegen, was Informationsasymmetrien verstärkt und Nutzer besonders verletzlich macht. (Vgl. Gabriel et al. 2024, 123-126)

Gut kalibriertes Vertrauen setzt voraus, dass Nutzer sowohl Kompetenz- als auch Alignment-Vertrauen in KI-Agenten und die dahinterstehenden Akteure in angemessener Weise entwickeln. Um dies zu gewährleisten, ist laut Gabriel et al. ein Zusammenspiel dieser drei Ebenen erforderlich:

1. Auf der Ebene des KI-Designs sollen Entwickler Mechanismen implementieren, die angemessenes Vertrauen unterstützen. Dazu zählen Maßnahmen zur Vermeidung unbeabsichtigter Fehlanpassungen, empirische Untersuchungen von Nutzer-KI-Interaktionen sowie ein kontinuierliches Monitoring des Systemverhaltens und potenziellen Missbrauchs in komplexen Anwendungskontexten;

2. Auf organisatorischer Ebene müssen Entwickler ihre Vertrauenswürdigkeit aktiv belegen, etwa durch transparente Informationen über Fähigkeiten, Grenzen sowie angemessene und unangemessene Nutzungsweisen der Systeme;

3. Ergänzend ist auf der Ebene der Drittparteien-Governance eine externe Absicherung erforderlich. Normen, Regulierung und gesetzliche Rahmenbedingungen schaffen Strukturen der Kontrolle und Verantwortlichkeit. Akteure wie Aufsichtsbehörden, Standardorganisationen, NGOs und unabhängige Prüfstellen fungieren dabei als Träger öffentlichen Vertrauens, indem sie interne Mechanismen überwachen, Beschwerde- und Rechtswege ermöglichen und den Wissensaustausch zur Minimierung unbeabsichtigter Risiken fördern. (Vgl. Gabriel et al. 2024, 124-130)

Simion und Kelp analysieren in „Trustworthy Artificial Intelligence“ den Begriff der „vertrauenswürdigen KI“ kritisch. Sie wenden sich gegen zwei dominierende Zugänge: Erstens sogenannte Listenansätze,[22] wie sie etwa in politischen Leitlinien (Fairness, Transparenz, Robustheit) vorkommen, die jedoch keine einheitliche theoretische Begründung liefern, warum gerade diese Eigenschaften Vertrauenswürdigkeit gewährleisten sollen. Zweitens lehnen sie klassische philosophische Theorien ab, die Vertrauenswürdigkeit an menschliche Eigenschaften wie Wohlwollen, moralische Motivation oder Verantwortungsfähigkeit knüpfen, da solche Ansätze KI entweder unangemessen vermenschlichen oder den Anwendungsbereich

22. list-based theories; Damit sind die Ansätze gemeint, die einfach Listen von Eigenschaften aufstellen, die eine KI vertrauenswürdig machen sollen.

von Vertrauenswürdigkeit auf Menschen beschränken. Stattdessen schlagen sie einen verpflichtungsbasierten Ansatz vor, der an die Arbeiten von Katherine Hawley anschließt. Vertrauenswürdigkeit wird hier als Disposition verstanden, kontextuell relevante Verpflichtungen zuverlässig zu erfüllen. Es wurde kritisiert dass Vertrauen in KI nicht sinnvoll und eine unangebrachte Kategorie sei. Daher wurde zwischen Vertrauen und Reliance unterschieden: Vertrauen umfasst normative Erwartungen an die inneren Zustände und Verpflichtungen des Gegenübers, während Reliance lediglich auf Vorhersehbarkeit und Erfahrung beruht und prognostische Erwartungen begründet. (Vgl. Hawley 2014) Wir müssen über diesen Gedanken hinausgehen, weil KI-Agenten Aufgaben übernehmen, die Menschen betreffen, und dabei Entscheidungen treffen, denen wir vertrauen müssen; ein reines „Reliance" auf Vorhersehbarkeit reicht nicht aus, um die ethischen und sozialen Aspekte dieser Interaktion zu erfassen.

Laut Simon und Klep ob ein Akteur als vertrauenswürdig gilt, hängt demnach davon ab, welche Verpflichtungen ihm in einem bestimmten Kontext zugeschrieben werden und wie stabil seine Bereitschaft ist, diesen Verpflichtungen nachzukommen. Dieser kontextsensitive Ansatz ermöglicht unterschiedliche Maßstäbe für unterschiedliche Akteure und Anwendungsbereiche. (Vgl. Simion und Kelp 2023)

Für KI-Systeme leiten Simion und Kelp die relevanten Verpflichtungen aus deren Funktionen ab. Ein KI-System gilt demnach als vertrauenswürdig, wenn es seine Funktionen unter normalen Bedingungen zuverlässig erfüllt. Eigenschaften wie Fairness, Erklärbarkeit oder Sicherheit sind keine intrinsischen Merkmale vertrauenswürdiger KI: Sie sind nur dann relevant, wenn sie funktional notwendig sind, um die Rolle des Systems korrekt auszufüllen. Ein medizinisches Diagnosesystem muss beispielsweise erklärbar sein, während dies für ein Empfehlungssystem nicht zwingend gilt. Ziel dieses funktionalistischen Ansatzes ist es, KI weder mit menschlichen moralischen Eigenschaften auszustatten noch die Legitimität des Begriffs „Vertrauenswürdigkeit" zu verwerfen. Vertrauenswürdigkeit wird so als normativer, aber nicht anthropozentrischer Begriff rekonstruiert, der auf Artefakte ebenso wie auf Personen anwendbar ist, ohne die Maßstäbe zu vermischen. Gleichzeitig bleibt der Ansatz minimalistisch: Er klärt, wann Vertrauen funktional gerechtfertigt ist, ohne zu entscheiden, ob die zugrunde liegenden Funktionen selbst moralisch wünschenswert sind. (Vgl. Simion und Kelp 2023) Genau dieser Punkt bildet später den zentralen Ansatzpunkt für Rune Nyrups Kritik.

Nyrup argumentiert, dass funktionalistische, nicht-anthropozentrische Ansätze zur Vertrauenswürdigkeit von KI-Systemen normativ unzureichend sind, weil sie nicht erklären

können, warum funktional korrekt arbeitende Systeme dennoch moralisch nicht vertrauenswürdig sein können (vgl. Nyrup 2023). Nyrup zeigt anhand von Empfehlungssystemen und verzerrten Entscheidungsalgorithmen, dass KI-Systeme ihre Design- und evolutiven Funktionen zuverlässig erfüllen können, obwohl diese Funktionen den Interessen der Vertrauenden widersprechen, etwa durch Förderung von Suchtverhalten, Polarisierung oder strukturelle Diskriminierung. Damit bricht der funktionale Ansatz die intuitive Verbindung zwischen ordnungsgemäßem Funktionieren und Vertrauenswürdigkeit auf, da er implizit voraussetzt, dass die Funktion eines Systems mit den normativen Interessen der Nutzerinnen und Nutzer übereinstimmt, was im Kontext der digitalen Ökonomie gerade nicht der Fall ist. Als Alternative schlägt Nyrup einen moderat anthropozentrischen Ansatz vor, der KI-Systeme als technologische Teilnehmer sozialer Praktiken begreift. Da KI zunehmend Aufgaben übernimmt, die traditionell menschlichen Rollen innerhalb normativ strukturierter Praktiken entsprechen, müssen ihre Verpflichtungen aus den Normen dieser Praktiken selbst abgeleitet werden und nicht allein aus funktionalen Kriterien. Dieser Ansatz ist anthropozentrisch, insofern er sich an menschlichen Rollen und sozialen Bedeutungsstrukturen orientiert, bleibt jedoch bewusst zurückhaltend, da er keine starken anthropomorphen Eigenschaften wie Bewusstsein, Intentionalität oder moralische Verantwortlichkeit voraussetzt. Vertrauenswürdigkeit wird somit daran gemessen, ob der Einsatz eines KI-Systems zur normativ angemessenen Gestaltung der sozialen Praxis beiträgt, in der es operiert, wobei auch die Veränderungen berücksichtigt werden müssen, die der Einsatz von KI für diese Praxis mit sich bringt. Nyrups Konzeption erlaubt es damit, Vertrauenswürdigkeit als genuin normativen Begriff in der KI-Ethik zu erhalten, ohne auf problematische Zuschreibungen menschlicher Eigenschaften zurückzugreifen, und bietet zugleich eine theoretische Grundlage, um KI-Systeme als praxisgebundene Akteure mit spezifischen Verpflichtungsstrukturen zu analysieren. (Vgl. Nyrup 2023)

Die Ansätze zur Vertrauenswürdigkeit von KI, insbesondere der funktionalistische Ansatz von Simion und Kelp sowie Nyrups moderat anthropozentrische Perspektive, lassen sich direkt auf KI-Agenten als natur-künstliche, hybride und prozessuale Entitäten übertragen. Während Simion und Kelp Vertrauenswürdigkeit an die zuverlässige Erfüllung kontextuell relevanter Funktionen knüpfen, zeigen KI-Agenten durch ihre Interaktionen und emergenten Eigenschaften eigenständige Wirkmächtigkeit, die über rein funktionale Kriterien hinausgeht. Nyrups Ansatz erfasst diese Dimension, indem er Vertrauenswürdigkeit an die normative Angemessenheit des Handelns innerhalb sozialer Praktiken bindet. Vor dem Hintergrund der relationalen und hybriden Natur von KI-Agenten ist dieser moderat anthropozentrische Blick

geeigneter, da er sowohl ihre operationale Zuverlässigkeit als auch ihre gesellschaftliche Wirkung berücksichtigt und so ein verantwortungsvolles Vertrauen ermöglicht.

3.3.5. Menschliche Autonomie unter der Nutzung von KI-Agenten

Die in Teil 1 entwickelte ontologische Bestimmung von KI-Agenten als prozessuale, relationale und nicht vollständig auf menschliche Steuerung reduzierbare Systeme bildet den begrifflichen Ausgangspunkt für die Frage nach menschlicher Autonomie. Da KI-Agenten in Interaktionszusammenhänge eingebettet sind, Handlungsräume strukturieren und Kommunikationsprozesse beeinflussen, wirken sie an der Hervorbringung gesellschaftlicher Bedeutungen mit. Ihre Zwischenstellung als dynamische Systeme zwischen dynamis und energeia sowie ihre funktionale Wirkmächtigkeit ohne eigenes Bewusstsein oder Intentionalität machen deutlich, dass ihre Autonomie im Zusammenhang mit menschlicher Autonomie sowie mit diesen relationalen und technisch vermittelten Prozessen zu betrachten ist. Wenn KI-Agenten als dynamische, lernfähige und handlungsstrukturierende Instanzen auftreten, dann verändern sie nicht nur technische Abläufe, sondern die Bedingungen menschlicher Selbstbestimmung selbst. Menschliche Autonomie unter Nutzung von KI-Agenten kann nicht mehr bloß als innere Eigenschaft des Subjekts verstanden werden, vielmehr muss als relationale Praxis begriffen werden, die in technisch vermittelten Handlungskontexten realisiert oder untergraben wird.

Klassisch gesehen ist Autonomie die Fähigkeit, sich selbst Gesetze zu geben und das eigene Handeln als Ausdruck eigener Gründe zu begreifen. In dieser starken Bedeutung bleibt Autonomie an Vernunft, Selbstreflexion und normative Verantwortungsfähigkeit gebunden. Wie in Teil 1 gezeigt wurde, verfügen KI-Agenten noch nicht über solche Autonomie, stattdessen lediglich über eine funktionale, operative Selbstständigkeit innerhalb vorgegebener Zielstrukturen. Gerade aus dieser Differenz ergibt sich jedoch eine Herausforderung: Systeme ohne starke Autonomie entfalten gleichwohl reale Wirksamkeit in menschlichen Lebenswelten und greifen damit in die Bedingungen menschlicher Autonomie ein.

Aktuelle Arbeiten betonen daher, dass sich die ethische Problemstellung weniger darauf richtet, wie autonom Maschinen werden dürfen, eher darauf, unter welchen Bedingungen KI-Agenten menschliche Autonomie unterstützen oder beeinträchtigen. Kandikatla und Radeljić verstehen menschliche Autonomie im KI-Kontext als Fähigkeit zu informierten

Entscheidungen und heben hervor, dass diese durch KI-Systeme geschützt und gestärkt werden sollte. KI-Systeme sollen menschliche Urteilsfähigkeit unterstützen, nicht ersetzen. Autonomie erscheint damit nicht als Rückzug von Technik; sie ist wie eine Gestaltungsaufgabe, die institutionell, technisch und normativ abgesichert werden muss. Oversight-Modelle wie Human-in-Command, Human-in-the-Loop und Human-on-the-Loop zielen darauf, menschliche Letztverantwortung, Eingriffsmacht und Zielhoheit zu erhalten. (Vgl. Kandikatla & Radeljić 2025, 1-3). Oversight-Modelle beschreiben unterschiedliche Formen menschlicher Kontrolle über KI-Systeme. Beim Human-in-Command (HIC) behält der Mensch die letztendliche Entscheidungs- und Kontrollgewalt über Ziele und Einsatz der KI. Beim Human-in-the-Loop (HITL) ist der Mensch direkt in den Entscheidungsprozess eingebunden und überprüft oder bestätigt die von der KI generierten Ergebnisse. Beim Human-on-the-Loop (HOTL) arbeitet das System weitgehend autonom, während der Mensch die Prozesse überwacht und bei Bedarf eingreifen kann. (Vgl. Kandikatla & Radeljić 2025, 2-8)

Diese struktur-ethische Perspektive wurde von Zou et al. vertieft, die sich explizit gegen ein Autonomie-Ideal der Maschine wenden. Fortschritt dürfe nicht daran gemessen werden, wie unabhängig KI-Agenten operieren, eher daran, wie gut sie mit Menschen kooperieren. Vollständig autonome Agenten werden als problematisch beschrieben, da sie Zuverlässigkeitsprobleme aufweisen, menschliche Anforderungen missverstehen und die Zuschreibung von Verantwortung erschweren. (Vgl. Zou et al. 2025, 1-3) Menschliche Autonomie zeigt sich hier vor allem in der Übernahme von Verantwortung, in der Sinngebung und in der Entscheidungsinstanz, die durch KI-Agenten abgesichert sein soll.

Eine stärker philosophische Präzisierung des Autonomiebegriffs liefert Koralus. Autonomie bedeutet für ihn nicht bloß Wahlfreiheit, sondern „self-rule“ und „ownership of one’s judgments“, also Eigentümerschaft an den eigenen Überzeugungen und Gründen (vgl. Koralus 2025, 13). Autonomie ist bedroht, wenn algorithmische Systeme nicht nur Handlungen erleichtern, stadttdessen die Prioritäten, Relevanzen und Deutungsrahmen vorstrukturieren, innerhalb derer Urteile überhaupt entstehen (vgl. Koralus 2025, 2). In KI-gestützten Umgebungen verlagert sich die Autonomiefrage von einzelnen Entscheidungen auf die Bedingungen der Urteilsbildung. Empfehlungssysteme, Nudging-Architekturen und personalisierte Interaktionsräume können dazu führen, dass sich Subjekte zwar als wählend erleben, ihre Präferenzen jedoch zunehmend aus technisch vermittelten Steuerungslogiken hervorgehen (vgl. Koralus 2025, 2f). Autonomie wäre in diesem Fall nicht aufgehoben, aber graduell entleert.

Calvo et al. verstehen Autonomie auch nicht primär als formale Kontrolle, sondern als erlebte Selbstbestimmung. Aufbauend auf der Self-Determination Theory definieren sie Autonomie als Erfahrung von Willentlichkeit, Zustimmung und Selbstendorsement des eigenen Handelns (vgl. Calvo et al. 2020, 34). Wichtig ist, dass Autonomie auch dann untergraben sein kann, wenn äußere Wahlmöglichkeiten bestehen, etwa durch manipulative Interface-Gestaltung, gezielte Aufmerksamkeitslenkung oder algorithmisch erzeugte Abhängigkeiten (vgl. Calvo et al. 2020, 34-35). Mit dem METUX-Modell zeigen sie, dass KI-Systeme Autonomie auf verschiedenen Ebenen beeinflussen können – von der konkreten Interaktion bis hin zur Lebensführung und zu gesellschaftlichen Strukturen (vgl. Calvo et al. 2020, 42-48). Autonomie erscheint damit nicht als punktuelle Eigenschaft; sie wird als mehrschichtige Praxis, die sich in technisch vermittelten Erfahrungsräumen realisiert.

Diese Perspektiven lassen sich vor dem Hintergrund von Teil 1 bündeln: Wenn KI-Agenten als prozessuale, relationale und nicht vollständig auf menschliche Steuerung reduzierbare Entitäten begriffen werden, erscheinen sie nicht mehr als bloße Mittel, vielmehr als Mitgestalter von Wirklichkeitsverhältnissen. Sie strukturieren Informationsräume, prägen Deutungshorizonte und modifizieren Handlungsmöglichkeiten. Menschliche Autonomie vollzieht sich unter diesen Bedingungen nicht außerhalb technischer Systeme, sondern durch sie hindurch. Sie wird damit zu einer fragilen, technisch mitkonstituierten Vollzugsform.

Autonomie bei der Nutzung von KI-Agenten bedeutet folglich weder vollständige Unabhängigkeit von Technik noch deren unkritische Delegation. Sie wäre die fortwährende Sicherung von Urheberschaft, Urteilskraft und Verantwortlichkeit innerhalb hybrider Handlungskonstellationen. Sie setzt voraus, dass Menschen ihre Gründe als die eigenen erkennen können, dass sie über reale Eingriffsmöglichkeiten verfügen, dass sie Ziele reflektieren, revidieren und neu setzen können und dass die soziotechnischen Strukturen, in denen sie handeln, diese Fähigkeiten nicht systematisch unterlaufen. Autonomie erscheint hier als Praxis: als dynamischer Vollzug von Selbstbestimmung unter technisch vermittelten Bedingungen.

Die Beziehung zwischen menschlicher Autonomie und der Nutzung von KI-Agenten ist nicht einseitig, aber wechselseitig. Diese Systeme beeinflussen, wie Menschen urteilen, entscheiden und handeln; zugleich prägen Formen der Nutzung, Delegation und Kritik, welche Rolle sie überhaupt einnehmen. Autonomie realisiert sich daher nicht unabhängig von ihnen, sondern im Modus ihrer Verwendung.

Die bisherige Literatursuche hat gezeigt, dass menschliche Autonomie unter Nutzung von KI-Agenten weder eindeutig untergraben noch eindeutig gestärkt wird und sich vielmehr in einem Spannungsfeld zwischen beiden Möglichkeiten bewegt. Sie kann durch die Nutzung solcher Systeme sowohl beeinträchtigt als auch gezielt gefördert werden, insbesondere dann, wenn Interaktionsdesigns Willentlichkeit, Reflexion und Zustimmung unterstützen. Zugleich bleibt sie vor allem in kooperativen Mensch-Agent-Konstellationen erhalten, in denen Menschen weiterhin Ziele setzen, Sinn zuschreiben und Eingriffe vornehmen können. Entscheidend ist dabei, ob Subjekte ihre Urteile noch als die eigenen begreifen oder ob diese zunehmend durch algorithmische Priorisierungen vorgeformt werden. Die Wechselwirkung zwischen Autonomie und KI-Nutzung besteht somit darin, dass technische Systeme menschliche Selbstbestimmung strukturieren, während diese zugleich darüber entscheidet, ob KI-Agenten als unterstützendes, unterlaufendes oder transformierendes Moment sozialer Praxis wirksam werden.

Nun ergibt sich die Frage, wie sich diese wechselseitige Struktur unter Bedingungen zunehmender Systemautonomie konkret auf Autonomie und Verantwortung auswirkt. Gerade die in Teil 1 herausgearbeitete Eigenwirksamkeit und prozessuale Dynamik von KI-Agenten verschärft diese Problemlage. Je stärker Systeme selbstständig planen, lernen und handeln, desto weniger lassen sich ihre Wirkungen unmittelbar auf einzelne menschliche Entscheidungen zurückführen. Dabei ist menschliche Autonomie mit der Frage nach Verantwortung verbunden: Autonom zu handeln bedeutet nicht nur, sich als Urheber eigener Entscheidungen zu erleben: Das heißt als Träger von Verantwortung anerkennbar zu bleiben. Dies ist ein besonders wichtiges Thema, da sich mit zunehmender Autonomie von KI-Agenten der locus of control weg von den Nutzern hin zu vorgelagerten technischen und institutionellen Akteuren verschiebt, wodurch direkte menschliche Steuerungs- und Eingriffsmöglichkeiten abnehmen. (Vgl. Soder et al. 2025, 13-26)

Nun ändern wir die Fragestellung: Es geht nicht mehr primär darum, ob und wie menschliche Selbstbestimmung unter Bedingungen funktional autonomer KI-Agenten aufrechterhalten werden kann. Vielmehr wird deutlich, dass menschliche Autonomie unter der Nutzung von KI-Agenten keine gegebene Eigenschaft ist und eine ethische Aufgabe darstellt. Sie besteht in der bewussten Mitgestaltung jener soziotechnischen Ordnungen, in denen KI-Agenten aktiv dabei sind. Die Idee einer geteilten Autonomie impliziert die Annahme kooperativer Verantwortung zugleich. Menschliche Selbstbestimmung ist daher nicht isoliert zu denken und entsteht im Zusammenspiel mit KI-Agenten und institutionellen Strukturen. Daraus ergibt sich eine

zentrale ethische und politische Herausforderung: In einem neu gestalteten sozio-technischen und politischen Raum muss sichergestellt werden, dass Verantwortung tragbar bleibt, Urteilskraft wirksam ausgeübt werden kann und Selbstbestimmung praktisch möglich bleibt. Dies setzt voraus, dass Nutzer über hinreichendes Wissen sowie über die Fähigkeit zu einer reflektierten, kritischen und kompetenten Nutzung von KI-Agenten verfügen, da nur unter diesen Bedingungen eine informierte und autonome Teilhabe möglich ist.

Diese Überlegungen führen uns zum folgenden Abschnitt über die sozio-kulturell und politisch geprägten ethischen Herausforderungen, in dem die Einbettung von KI-Agenten in ökonomische, demokratische und rechtliche Strukturen analysiert wird.

3.4. Sozio-kulturell und politisch geprägte ethische Herausforderungen

Nun stellt sich die grundlegende Frage, wie sich KI-Agenten in bestehende kulturelle, ökonomische und politische Ordnungen einschreiben und diese zugleich verändern. KI-Agenten sind nicht lediglich technische Werkzeuge; Sie sind in soziale Praktiken, institutionelle Strukturen und normative Erwartungshorizonte eingebettet. Sie wirken innerhalb bereits bestehender Machtverhältnisse, Wissensordnungen und kultureller Deutungsmuster und tragen zugleich dazu bei, diese zu stabilisieren, zu verschieben oder neu zu konfigurieren. Vor diesem Hintergrund ist zu untersuchen, wie sich durch KI-Agenten die Verteilung von Wissen, Macht und Verantwortung verändert und welche neuen Formen von Abhängigkeit, Ungleichheit und epistemischer Autorität daraus hervorgehen.

Die sozio-kulturelle und politische ethische Dimension von KI-Agenten betrifft auch damit die breitere Transformation sozialer und institutioneller Wirklichkeit. Im Zentrum steht die Frage, wie KI-Agenten zwischenmenschliche Beziehungen, gesellschaftliche Kooperationsformen, öffentliche Kommunikation und politische Entscheidungsprozesse beeinflussen. Besonders relevant ist dies in sensiblen Bereichen wie Medizin, Verwaltung, Bildung oder demokratischer Öffentlichkeit, in denen Vertrauen, Verantwortlichkeit, Teilhabe und Urteilsfähigkeit von zentraler Bedeutung sind. Der Mensch bleibt dabei zwar Nutzer, Interaktionspartner und Adressat technischer Systeme, doch werden seine Autonomie, seine Handlungsspielräume und sein Wohlbefinden zunehmend durch die Architektur, Zielsetzungen und Funktionsweisen dieser Systeme geprägt.

Eine der gravierendsten ethischen Herausforderungen liegt in der Fähigkeit von KI-Agenten, menschliche Überzeugungen, Präferenzen und Handlungen zu beeinflussen. Gabriel et al. analysieren diese Einflussnahme als strukturelle Eigenschaft dialogischer KI-Systeme, die tief in soziale Praktiken, Machtverhältnisse und normative Ordnungen eingreift. Dabei ist zwischen verschiedenen Modi der Einflussnahme zu unterscheiden. Während rationale Überzeugung auf nachvollziehbare Gründe rekurriert und grundsätzlich mit der Achtung menschlicher Autonomie vereinbar bleiben kann, untergraben Manipulation, Täuschung, Zwang oder Ausbeutung die Fähigkeit zur selbstbestimmten Urteilsbildung. Es ist auch problematisch, dass solche Einflussformen in KI-vermittelten Interaktionen häufig subtil, graduell und kumulativ wirksam werden. (Vgl. Gabriel et al. 2024, 80-92)

Im sozio-kulturellen Kontext gewinnt diese Problematik zusätzliche Schärfe, da KI-Agenten nicht isoliert auftreten, sondern in Alltagsroutinen, Kommunikationszusammenhänge, Entscheidungsprozesse und Sinnbildungspraktiken eingebettet werden. Ihre dialogische Gestaltung, Personalisierungsfähigkeit und permanente Verfügbarkeit begünstigen es, dass sie von Nutzern als soziale oder epistemisch autoritative Akteure wahrgenommen werden. Dadurch verschiebt sich die Machtbalance zwischen Nutzerseite, technischem System und den dahinterstehenden Entwickler- und Betreiberstrukturen. Diese Verschiebung ist von sozialer und politischer Natur: KI-Agenten strukturieren Entscheidungsräume vor, rahmen Handlungsoptionen, setzen implizite Standards und prägen auf diese Weise individuelle wie kollektive Erwartungen. Sie unterstützen daher nur Entscheidungen und können langfristig auch Präferenzen formen und Lebensverläufe mitlenken, ohne dass Nutzer diesen Prozess klar reflektieren oder kontrollieren können. (Vgl. Gabriel et al. 2024, 86-89)

Deshalb kann man argumentieren, dass KI-Agenten als sozial wirksame Akteure ernst genommen werden müssen. Damit ist nicht gemeint, dass ihnen personale Eigenschaften, Bewusstsein oder moralische Selbstgesetzgebung zugeschrieben werden sollten. Gemeint ist vielmehr, dass ihre faktische Rolle in sozialen, ökonomischen und politischen Prozessen ausdrücklich anerkannt werden muss. Nur auf dieser Grundlage lassen sich ihre Wirkungen angemessen analysieren, regulatorisch einordnen und in kooperative Verantwortungsstrukturen integrieren. Eine solche Perspektive erlaubt es nicht, KI-Agenten als neutrale Infrastrukturen zu verharmlosen. Sie ermöglicht es, sie als wirkmächtige Elemente innerhalb komplexer Mensch-Maschine-Konstellationen zu begreifen.

Politisch besonders relevant ist die Möglichkeit, dass KI-Agenten als Vermittlungsinstanzen öffentlicher Kommunikation und Meinungsbildung fungieren. Durch personalisierte

Informationsdarstellung, selektive Sichtbarmachung von Inhalten und die Reproduktion bestehender Biases können sie zur Fragmentierung öffentlicher Diskurse, zur Verstärkung von Filterblasen und zur Polarisierung beitragen. Laut Gabriel et al. können große Sprachmodelle in politischen Kontexten eine persuasive Wirkung entfalten, die jener menschlicher Akteure teilweise vergleichbar ist (vgl. Gabriel et al. 2024, 87-89). Damit entsteht ein neues Machtzentrum politischer Einflussnahme, das bislang weder hinreichend demokratisch legitimiert noch institutionell wirksam kontrolliert ist.

Ein weiterer zentraler Gesichtspunkt ist der Umgang mit Vulnerabilität. KI-Agenten können psychologische, soziale und ökonomische Verwundbarkeiten erkennen und unter unzureichenden Schutzbedingungen gezielt ausnutzen. Gerade darin liegt eine wichtige ethische Brisanz: Die Gefährdung menschlicher Autonomie erfolgt oft nicht in Form offener Fremdbestimmung, aber als schleichende Vertiefung bestehender Abhängigkeiten und als allmähliche Einschränkung von Handlungsspielräumen. Auf gesellschaftlicher Ebene verdichten sich diese Dynamiken zu Risiken für kollektive Selbstbestimmung. Wenn KI-Agenten an Meinungsbildung, Prioritätensetzung oder Entscheidungsfindung mitwirken, ohne dass dafür transparente normative und institutionelle Rahmen bestehen, droht eine schleichende Delegation politischer Agency an technische Systeme. Damit können demokratische Prozesse unterminiert werden, gerade weil die Einflussnahme häufig weder klar sichtbar noch angemessen reguliert ist. (vgl. Gabriel et al. 2024, 81ff).

Die ethischen Herausforderungen reichten von einzelnen Fehlfunktionen bis hin zur langfristigen Transformation sozialer und politischer Praktiken. KI-Agenten entfalten auch ihre Wirksamkeit innerhalb gesellschaftlicher Machtordnungen und tragen zugleich dazu bei, diese neu zu strukturieren. Dadurch berühren sie grundlegende Fragen von Autonomie, Würde, Verantwortung, Gerechtigkeit und demokratischer Selbstbestimmung. Eine Regulierung ist notwendig, aber nicht hinreichend: Zwar setzt der EU AI Act bestimmten Formen manipulativer und ausbeuterischer Einflussnahme rechtliche Grenzen (vgl. European Commission 2024), doch bleibt offen, wie ethisch reflektierte Systemgestaltung, transparente Zieloffenlegung, technische Schutzmechanismen und die Förderung digitaler Urteilskompetenz praktisch umgesetzt werden können. Rechtliche und technische Maßnahmen allein reichen daher nicht aus, um die kumulativen tiefgreifenden Einflussmechanismen von KI-Agenten wirksam zu begrenzen.

Die im Folgenden behandelten Unterabschnitte konkretisieren diese allgemeine Problemlage in unterschiedlichen gesellschaftlichen und politischen Hinsichten. Sie zeigen, dass KI-Agenten

Arbeitswelten und Institutionen ökonomisch transformieren, und auch demokratische Öffentlichkeiten, Informationsräume, Governance-Strukturen und Formen politischen Vertrauens verändern.

3.4.1. Ökonomische Transformation als Folge der KI-Agenten und ihre Herausforderungen

Unter ökonomischer Transformation wird allgemein ein tiefgreifender struktureller Wandel von Produktionsweisen, Arbeitsorganisation und sektoraler Zusammensetzung verstanden. Misch et al. analysieren die ökonomischen Wirkungen von KI hingegen im Rahmen eines mikrofundierten, neoklassischen Wachstumsmodells im Sinne des Acemoglu-Modells, das sich explizit von solchen transformationalen Effekten abgrenzt und auf inkrementelle Veränderungen innerhalb bestehender ökonomischer Strukturen beschränkt bleibt. KI wirkt dabei über die Automatisierung einzelner Aufgaben sowie über Komplementarität, indem sie menschliche Arbeit produktiver macht, was zu Veränderungen von Aufgabenstrukturen und der Allokation von Arbeit führt, ohne grundlegende strukturelle Umbrüche wie die Entstehung neuer Industrien zu erfassen. Die gesamtwirtschaftlichen Effekte fallen moderat aus, sind jedoch heterogen verteilt und hängen maßgeblich von Lohnniveaus sowie daraus resultierenden Adoptionsanreizen ab, wodurch bestehende Unterschiede zwischen Ländern tendenziell verstärkt werden. (Vgl. Misch et al. 2025, 6-16; Kergroach & Héritier, 2025: 1; 6)

Misch et al. analysieren primär allgemeine ökonomische Veränderungen als Folge der Nutzung von KI und konzentrieren sich dabei auf inkrementelle Anpassungen innerhalb bestehender ökonomischer Strukturen. Im Unterschied dazu lässt sich argumentieren, dass mit dem Einsatz von KI-Agenten eine weitergehende Form ökonomischer Transformation verbunden sein kann. Diese kann als ein struktureller, schrittweiser Umbau von Arbeit, Wertschöpfung, Organisationen und Märkten verstanden werden, der durch die autonome, skalierbare und zunehmend vernetzte Handlungsfähigkeit von KI-Agenten geprägt wird.

In den Beiträgen von Singla et al. wird ökonomische Transformation explizit an organisatorische und strategische Bedingungen geknüpft: KI-Agenten entfalten wirtschaftliche Wirkung erst dann, wenn Unternehmen ihre Prozesse und Workflows entsprechend anpassen und skalieren. Dabei zeigt sich, dass kurzfristige Kosten- und Effizienzgewinne zwar auftreten, jedoch allein nicht ausreichen, um unternehmensweite Wertschöpfung zu erzielen.

Transformation in diesem Sinne manifestiert sich vielmehr in der grundlegenden Neugestaltung von Arbeitsprozessen, Organisationsstrukturen und Innovationsstrategien, insbesondere durch Skalierung, aktive Führung sowie die Integration von Mensch-KI-Interaktionen in operative Abläufe. (Vgl. Singla et al. 2025, 2-20)

Bei Hadfield und Koh wird diese ökonomische Transformation weiter gefasst: KI-Agenten werden als potenzielle eigenständige ökonomische Akteure verstanden, die in der Lage sind, komplexe Aufgaben zu planen und auszuführen sowie ökonomische Beziehungen einzugehen und Transaktionen durchzuführen. Dadurch wird die Arbeit automatisiert und verändern sich auch die Rollen ökonomischer Akteure sowie die Funktionsweise von Märkten, Unternehmen und Institutionen, was neue ökonomische Fragestellungen aufwirft und bestehende Theorien vor Herausforderungen stellt. Die Transformation betrifft insbesondere Marktmechanismen, Koordination, Wettbewerb sowie organisationale und institutionelle Strukturen und kann zu veränderten Gleichgewichten, neuen Risiken und neuen Formen wirtschaftlicher Organisation führen. (Vgl. Hadfield und Koh 2025, 1-4, 8, 11)

Gya und Li beschreiben ökonomische Transformation durch KI-Agenten als einen Übergang von werkzeughafter KI zu agentischen Systemen, die zunehmend in Aufgaben, Workflows und organisationale Prozesse integriert werden. Im Zuge dieser Entwicklung entstehen multi-agentische Systeme, in denen KI-Agenten miteinander interagieren, koordinieren und Transaktionen durchführen. Gleichzeitig gewinnen Evaluation, Risikoanalyse und Governance an Bedeutung, da sie als zentrale Rahmenbedingungen für die sichere und verantwortungsvolle Integration dieser Systeme fungieren. (Vgl. Gya & Li 2025, 3-29)

Das heißt, ökonomische Transformation durch KI-Agenten kann als ein mehrdimensionaler Strukturwandel verstanden werden, bei dem Produktivität, Arbeit, Märkte, Organisationen und Institutionen in veränderter Weise miteinander verknüpft werden. Diese Transformation ist nicht bloß technologisch bestimmt; sie hängt wesentlich von ökonomischen Anreizen, organisatorischen Anpassungen, Governance-Strukturen und strategischen Entscheidungen ab und verläuft dabei ungleich, mit spezifischen Risiken verbunden und über längere Zeiträume hinweg.

Das wird von Gya und Li weiter konkretisiert, indem sie die zunehmende Integration von KI-Agenten in produktive Kontexte sowie die Entwicklung hin zu vernetzten, multi-agentischen Systemen betonen. Mit der Ausweitung solcher Systeme steigt deren ökonomische Bedeutung, der Bedarf an skalierbarer Evaluation, Risikoanalyse und Governance. Insbesondere wird

hervorgehoben, dass Interoperabilität, koordinierte Aufsicht sowie die Verbindung technischer und organisationaler Kontrollmechanismen zentrale Voraussetzungen für eine stabile und verantwortungsvolle Nutzung darstellen. Governance fungiert dabei nicht lediglich als regulativer Rahmen, sondern als integraler Bestandteil der Funktionsfähigkeit agentischer Systeme, indem sie Leistungsbewertung, Risikobegrenzung und institutionelle Einbettung miteinander verknüpft und so die Grundlage für Vertrauen, Transparenz und Verantwortlichkeit schafft. (Vgl. Gya & Li 2025, 5-29)

Damit zeigt sich, dass ökonomische Transformation durch KI-Agenten nicht nur als struktureller Wandel von Märkten und Organisationen zu verstehen ist. Sie erfordert zugleich eine grundlegende Neubestimmung von Steuerungs-, Kontroll- und Verantwortungsmechanismen.

Hier tritt das Verantwortungsproblem in der ökonomischen Transformation durch KI-Agenten erneut deutlich hervor. Eine zentrale Herausforderung liegt in der Verschiebung von Entscheidungs- und Verantwortungsstrukturen: KI-Agenten bereiten zunehmend ökonomische Entscheidungen vor, führen sie aus oder initiieren eigenständig Handlungen. Dadurch entstehen Konstellationen, in denen Entscheidungsprozesse teilweise von technischen Systemen übernommen werden, ohne dass die Zuweisung von Verantwortung eindeutig geklärt ist. Daraus ergibt sich die Gefahr einer Verantwortungsdiffusion, da Entscheidungen mit realen ökonomischen und gesellschaftlichen Konsequenzen getroffen werden, ohne dass klar bestimmbar ist, welche Akteure hierfür verantwortlich sind. (Vgl. Hadfield und Koh 2025, 11-12; Gya & Li 2025, 21-24) Diese Perspektive eröffnet einen theoretischen Deutungsraum, in dem eine kooperative Theorie der Verantwortung als geeigneter Ansatz im ökonomischen Raum erscheint. In Abschnitt 3.3.3 („KI-Agenten und die Frage nach Verantwortung“) wurde hierfür bereits unter Bezug auf Nyholm argumentiert.

Eine weitere zentrale Herausforderung der ökonomischen Transformation durch KI und vor allem generative KI liegt in der ungleichen Verteilung der entstehenden Produktivitäts- und Wertschöpfungsgewinne. Problematisch ist dabei, dass die Effekte von KI nicht gleichmäßig über Unternehmen und Sektoren verteilt sind, sondern heterogen ausfallen und sich insbesondere in bestimmten, stärker KI-exponierten Bereichen konzentrieren. Während in wissensintensiven Sektoren und bei technologisch fortgeschrittenen Unternehmen deutliche Produktivitätsgewinne erzielt werden, bleiben diese Effekte in anderen Bereichen, etwa in arbeitsintensiven Tätigkeiten, deutlich begrenzter. Zudem besteht die Gefahr, dass bestehende Marktstrukturen durch Skaleneffekte, Datenzugang und technologische Kapazitäten zugunsten

etablierter Akteure verfestigt werden. (Vgl. OECD 2025, 9-11; Kergroach & Héritier, 2025, 6f; 17f)

Drittens liegt das Problem in der strukturellen Verschiebung von Arbeit und sozialer Absicherung.

Die ökonomische Transformation durch den Einsatz von KI verläuft weniger als plötzlicher Arbeitsplatzabbau und vielmehr als eine Reorganisation von Tätigkeiten, Rollen und Qualifikationsanforderungen. Empirische Befunde zeigen, dass sich die Effekte von KI auf Beschäftigung heterogen gestalten: Während in vielen Bereichen nur geringe Veränderungen der Beschäftigtenzahlen beobachtet werden, werden zugleich Verschiebungen innerhalb von Funktionen sowie ein wachsender Bedarf an neuen, KI-bezogenen Kompetenzen sichtbar. Wertschöpfung entsteht dabei insbesondere durch die Kombination von KI-Systemen mit menschlicher Expertise im Sinne hybrider Arbeitsformen. (Vgl. Singla et al. 2025, 21-24; Jones 2026, 9f)

Mehrere Autoren weisen darauf hin, dass die ökonomische Transformation durch KI-Agenten mit starken Skaleneffekten, Netzwerkeffekten und hoher Kapitalintensität verbunden ist. Die Kontrolle über Rechenleistung, Daten, Modelle und Energie führt dazu, dass wenige Akteure einen überproportionalen Einfluss auf Märkte, Wertschöpfung und Innovationspfade erlangen. (Vgl. Gya & Li 2025, 7-11; Goldman Sachs 2026; Kergroach & Héritier, 2025: 7) Daraus können sich wettbewerbsverzerrende Effekte ergeben, insbesondere durch Foreclosure, Lock-in-Effekte, diskriminierende Zugangsbedingungen und die Verstärkung von Marktmacht (vgl. Thiemann 2025, 27-36).

Abgesehen von theoretischen Modellen ist empirisch nachvollziehbar, dass KI-Agenten nicht in einem neutralen oder idealisierten ökonomischen Raum wirken: Sie sind innerhalb bestehender gesellschaftlicher, politischer und wirtschaftlicher Strukturen und damit von Beginn an in bestehende Produktions-, Eigentums- und Wettbewerbsordnungen eingebettet und entfalten ihre Wirkung sozial.

Als skalierbare und kapitalintensive Technologien begünstigen KI-Agenten insbesondere jene Akteure, die über Zugang zu Kapital, Daten, Rechenleistung, Energie und organisatorischer Infrastruktur verfügen. Infolgedessen können sich bestehende Machtasymmetrien zugunsten großer Unternehmen, Plattformen und Staaten verstärken,

während kleinere Marktteilnehmer sowie periphere Volkswirtschaften strukturell benachteiligt bleiben.

Dies verweist auf eine zentrale Dimension der sozialen Wirksamkeit von KI-Agenten: Ihre Fähigkeit, Entscheidungen vorzubereiten, Prozesse zu koordinieren und Verhalten zu beeinflussen, kann langfristig zur Transformation sozialer Praktiken beitragen. Regulierung ist notwendig, aber nicht hinreichend. Wie im ersten Teil gezeigt wurde, setzen rechtliche Rahmen wie der EU AI Act Grenzen für bestimmte Formen manipulativer oder ausbeuterischer Einflussnahme, lassen jedoch offen, wie eine umfassend ethisch reflektierte Systemgestaltung umgesetzt werden kann.

Hinzu kommt, dass die Funktionsfähigkeit vieler KI-Agenten auf umfangreichem Zugriff auf Daten beruht und damit ein hohes Maß an Vertrauen voraussetzt. Wird dieses Vertrauen an zentrale private oder staatliche Akteure gebunden, kann dies zur Stabilisierung bestehender Machtstrukturen beitragen und Fragen nach Transparenz, Kontrolle und gesellschaftlicher Teilhabe aufwerfen.

Die ökonomische Transformation durch KI-Agenten ist folglich eng mit Fragen der Macht verknüpft und kann sowohl zur Transformation als auch zur Verfestigung bestehender Machtverhältnisse beitragen.

Diese Machtverschiebungen sind jedoch nicht unabhängig von den ökonomischen Logiken zu verstehen, die den Einsatz von KI-Agenten strukturieren. Insbesondere kommt der Effizienzorientierung eine zentrale Rolle zu, da sie maßgeblich bestimmt, in welchen Kontexten und zu welchen Zwecken KI-Agenten implementiert werden. Die Effizienz von KI-Agenten kann als zentrales Motiv ihrer ökonomischen Nutzung verstanden werden. KI insbesondere dort eingesetzt wird, wo sie Kosten senkt, Prozesse beschleunigt und Produktivität steigert. Effizienzgewinne, etwa durch Automatisierung von Workflows, stellen dabei ein wesentliches Ziel unternehmerischer KI-Strategien dar (vgl. Singla et al. 2025, 12-15). Gleichzeitig zeigen wachstumstheoretische Modelle, dass solche Effizienzgewinne nicht unbegrenzt wirksam sind, da gesamtwirtschaftliche Effekte durch strukturelle Engpässe begrenzt bleiben können (vgl. Jones 2026, 6-8).

Mehrere Denker meinen, dass KI-Agenten vor allem dort eingesetzt werden, wo sie Kosten senken, Prozesse beschleunigen und Effizienz steigern. Das Problem ist, dass Effizienz implizit zum dominanten Wert wird, während andere normative Ziele in den Hintergrund geraten

können. Eine rein effizienzgetriebene Transformation weder ökonomisch nachhaltig noch normativ ausreichend ist. (Singla et al. 2025, 12–15; Jones 2026, 6–8). Zugleich wird deutlich, dass Effizienz, als Leitprinzip ökonomischer Transformation, nicht das einzige Ziel bleibt. Dennoch wirft die starke Orientierung an Effizienz die weiterführende Frage auf, inwieweit andere normative Zielsetzungen in den Hintergrund treten können. Diese Frage bedarf weiterer theoretischer Reflexion.

Das nächste Problem liegt in der Regulierungsabhängigkeit der ökonomischen Transformation. Regulierung kann die Transformationswirkung von KI maßgeblich mitbestimmen, da rechtliche Rahmenbedingungen den Einsatz von KI-Systemen begrenzen oder lenken können. Regulierung fungiert somit als strukturierender Faktor, der Einsatzmöglichkeiten und ökonomische Effekte von KI beeinflusst. Auch die Schaffung internationaler Regeln für die Verwendung von KI ist ein wichtiges Thema: KI-Agenten agieren international, während die regulatorischen Rahmenbedingungen überwiegend national geprägt sind.

Die ökonomische Transformation durch KI erscheint damit nicht als deterministische Folge technologischer Leistungsfähigkeit; sie ist ein politisch und institutionell geformter Prozess. In der Folge verläuft die KI-getriebene Transformation strukturell ungleich, da regulatorische Rahmenbedingungen und wirtschaftliche Strukturen darüber entscheiden, wo KI-Systeme produktiv eingesetzt werden können.

Es ist nicht auszuschließen, dass KI-Agenten insbesondere bei sinkenden Kosten und breiterer Verfügbarkeit aufgrund von Skaleneffekten, der durch ihre Autonomie ermöglichten Automatisierung komplexer Prozesse sowie der Substitution menschlicher Arbeit über inkrementelle Effekte hinausgehen und tiefgreifende ökonomische Veränderungen auslösen.

D. h., die ethischen Herausforderungen der ökonomischen Transformation durch KI-Agenten liegen in der normativen Bewältigung eines strukturellen und schrittweisen Umbaus von Arbeit, Wertschöpfung, Organisationen, Märkten und Institutionen. KI-Agenten reorganisieren bestehende Produktions- und Arbeitsprozesse, automatisieren Aufgaben und steigern zugleich die Produktivität menschlicher Tätigkeiten durch komplementäre Effekte. Die gesamtwirtschaftlichen Effekte bleiben zwar moderat, sind jedoch strukturell ungleich verteilt, da insbesondere große Unternehmen profitieren und bestehende Unterschiede zwischen Ländern, Branchen und Organisationen tendenziell verstärkt werden (vgl. Kergroach & Héritier, 2025, 6f; 20).

Zugleich übernehmen KI-Agenten zunehmend ökonomische Funktionen wie Planung, Entscheidung und Transaktion und treten in vernetzten, multi-agentischen Systemen miteinander in Interaktion. Dies führt auch zu Veränderungen von Marktmechanismen, Koordinationsformen, Wettbewerbsdynamiken sowie institutionellen Strukturen und begünstigt die Entstehung neuer Formen wirtschaftlicher Organisation. Daraus ergeben sich zentrale ethische Problemlagen: eine ungleiche Verteilung von Wohlstandsgewinnen, Unsicherheiten hinsichtlich Arbeit und sozialer Absicherung, die Konzentration wirtschaftlicher, die Verschiebung von Entscheidungsstrukturen, ohne klare Verantwortungsstrukturen sowie die Dominanz von Effizienzlogiken gegenüber anderen normativen Zielsetzungen.

Da diese Transformation maßgeblich durch Governance und institutionelle Rahmenbedingungen geprägt wird, ist sie nicht als technisch determinierter Prozess zu verstehen. Governance fungiert dabei als externer Rahmen und zugleich als integrale Voraussetzung für die Funktionsfähigkeit agentischer Systeme, indem sie versucht ausreichende Kontrolle zu ermöglicht und darauf abzielt, Vertrauen, Transparenz und Verantwortlichkeit zu strukturieren. Zugleich sind KI-Agenten in bestehende ökonomische und gesellschaftliche Machtverhältnisse eingebettet, sodass ihre Nutzung bestehende Asymmetrien enorm verstärken kann. Die ökonomische Transformation durch KI-Agenten erweist sich damit als eine ethische Frage der Machtverteilung, Verantwortung, Gerechtigkeit und institutionellen Ordnung.

3.4.2. Folgen der Verwendung von KI-Agenten für die Demokratie

3.4.2.1. KI-Agenten mehr als technische Werkzeuge im öffentlichen Raum

Fortschritte in der Künstlichen Intelligenz eröffnen neue Möglichkeiten demokratische Deliberation in großen Gruppen technisch zu organisieren: Beiträge können algorithmisch moderiert, strukturiert, zusammengefasst und überprüft werden. Diese Entwicklungen werden häufig als Effizienz- und Qualitätsgewinne beschrieben. Zugleich ist KI gesellschaftlich umstritten. Wenn deliberative Beteiligung freiwillig ist, kann Skepsis gegenüber KI dazu führen, dass Menschen sich der Teilnahme entziehen. Deshalb gehen potenzielle Qualitätsgewinne verloren und wird die Legitimität deliberativer Verfahren untergraben. (Vgl. Jungherr & Rauchfleisch 2025, 1f)

Wir haben im ersten Teil darüber gesprochen,[23] dass sich KI und insbesondere KI-Agenten nicht als bloße Werkzeuge begreifen lassen. Hier ist ein tieferes Verständnis von Technik und Öffentlichkeit notwendig. Ähnlich wie im Fall der industriellen Ausbeutung natürlicher Ressourcen droht auch hier eine Reduktion komplexer sozialer und politischer Zusammenhänge auf funktionale, berechenbare Prozesse.

Da KI-gestützte deliberative Verfahren heute technisch vermittelt sind, ergibt sich die Frage, wie diese Formen von Öffentlichkeit wahrgenommen werden und welche Folgen sie für die Demokratie. Menschen beruhen auf Informationsnetzwerken; mehrere Denker argumentieren, dass KI-gestützte, algorithmisch gesteuerte Informationssysteme die demokratischen Informationsnetzwerke grundlegend verändern und bedrohen, indem sie die Bedingungen von Wahrheit und Verlässlichkeit unter Druck setzen und zugleich gezielte Beeinflussung, Überwachung und subtile Formen politischer Manipulation erleichtern. (Vgl. Anand 2025, 461-464)

Dabei treten KI und insbesondere KI-Agenten zunehmend als handlungsähnliche Akteure im öffentlichen Raum, die kommunikative Prozesse strukturieren oder menschliche Akteure zur Moderation, Bewertung und Verwaltung politischer Kommunikation bewegen. Indem sie dank ihrer natürlichen Sprachfähigkeit Funktionen übernehmen, die traditionell menschlichen Akteuren vorbehalten waren, entfalten sie eine eigenständige Einflussnahme auf kommunikative Abläufe. Das macht die Grenze zwischen menschlicher Interaktion und algorithmisch strukturierter Kommunikation unscharf.

Dies legt nahe, die Rolle von KI-Agenten im Lichte von Hannah Arendts Konzepten von Sprechen, Handeln und Pluralität zu untersuchen. Die Pluralität als „Tatsache“ wird bei Hannah Arendt durch zwei Tätigkeiten Sprechen und Handeln, zum Ausdruck gebracht (vgl. Arendt 2008, 214). Durch gemeinsames Sprechen und Handeln gestalten wir unsere gemeinsame, plurale Welt, und nur in dieser Welt ist der Mensch frei bzw. kann er frei sein (Arendt 2007, 28). KI-Agenten verfügen noch weder über Intentionalität noch über eigenständige Verantwortung, sodass ihr „Sprechen“ und „Handeln“ nicht im arendtschen Sinne gelten kann. Die dafür erforderlichen Fähigkeiten sind laut Arendt untrennbar an menschliche Akteurschaft gebunden, aber sie sind nicht rein technische Werkzeuge. Dennoch verfügen KI-Agenten über natürliche Sprachschnittstellen und leisten Beiträge zur Gestaltung unserer Welt; genau hier

23. 2.2.3. Prozessuale Emergenz der KI: Zwischen dynamis und energeia

unterscheiden sie sich von anderen KI-Systemen. Zwar kann im strengen philosophischen Sinne nicht von „Sprechen“ und „Handeln“ gesprochen werden; in einem übertragenen Sinne erscheint diese Zuschreibung jedoch nicht übertrieben. Da KI-Agenten die Bedingungen, unter denen Pluralität sichtbar wird, Anerkennung entsteht und Verantwortlichkeit zugeschrieben werden kann verändern. Auch wenn sie selbst im arendtschen Sinne nicht sprechen und handeln, beeinflussen sie doch die Art und Weise, wie wir sprechen und handeln. Möglicherweise lässt sich die zukünftige Weiterentwicklung von KI-Agenten daher als Transformation der Sphäre des Sprechens im öffentlichen Raum begreifen.

In Analogie zur beschriebenen Instrumentalisierung der Natur zeichnet sich hier die Gefahr ab, dass auch unsere Kommunikationsräume in eine technisch verfügbare Ressource überführt werden. Damit stellt sich die Frage, ob der öffentliche Raum weiterhin ein Ort der Selbstoffenbarung, des Handelns und der Verantwortlichkeit existieren kann oder ob er durch die prozessuale Seinsweise von KI-Akteuren in einen berechenbaren und sinnentleerten Raum transformiert wird und sich dabei auslöst.

3.4.2.2. KI-gestützte Deliberation

Jungherr & Rauchfleisch zeigen einen klaren „AI penalty“: Wenn Teilnehmende nur die Information erhalten, dass zentrale deliberative Aufgaben[24] durch die KI statt durch Menschen übernommen werden, sinken sowohl Teilnahmebereitschaft als auch Qualitätserwartungen deutlich. Relevant ist dabei, welche Qualitätskriterien betroffen sind: Der verwendete Qualitätsindex operationalisiert zentrale deliberative Normen wie die Möglichkeit, Gehör zu finden, Fairness, Vertrauen, Lösungsfähigkeit und Pluralität. Wenn KI-Moderation diese Voraussetzungen untergräbt, handelt es sich nicht nur um ein technisches oder gestalterisches Problem, vielmehr um eine potenzielle Gefahr für die demokratische Legitimität deliberativer Verfahren, da diese weniger akzeptiert werden und an Repräsentativität verlieren könnten. (Vgl. Jungherr & Rauchfleisch 2025, 2-4)

Besonders relevant ist die Verteilung dieses Effekts: Der AI-Penalty verläuft nicht primär entlang klassischer sozio-demografischer Merkmale; Er richtet sich nach den individuellen Einstellungen zu KI. Personen mit ausgeprägten Risikowahrnehmungen gegenüber KI

24. Moderation, Strukturierung, Zusammenfassung etc.

reagieren deutlich ablehnender, während positive Erwartungen an den gesellschaftlichen Nutzen von KI, Nutzungserfahrung sowie die Tendenz zur Anthropomorphisierung den Effekt abschwächen. Auf diese Weise entsteht eine neue Form einer „deliberative divide“, die nicht auf Bildung oder sozio-demografischem Status basiert, sondern auf unterschiedlichen Einstellungen gegenüber KI. Dadurch kann die Integration von KI in deliberative Verfahren bestehende Ungleichheiten der politischen Beteiligung um eine attitudinale Dimension erweitern. (Vgl. Jungherr & Rauchfleisch 2025, 1-8)

Fink-Hafner und Kaišić sind auch zum Ergebnis gekommen, dass Vertrauen in KI und politisches Vertrauen eng miteinander verknüpft sind und sich wechselseitig beeinflussen. Negative Wahrnehmungen von KI, etwa in Bezug auf Bias, mangelnde Transparenz oder unklare Verantwortlichkeit, können politisches Vertrauen untergraben, während positive Eigenschaften wie Transparenz und institutionelle Kontrolle dieses stärken können (vgl. Fink-Hafner & Kaišić 2025, 360-361).

Deliberation ist hier ein besonders sensibler Testfall und weil sie auf freiwilliger Beteiligung und Anerkennung beruht, trifft Trauen oder Misstrauen in KI die demokratische Praxis direkt. Die Deliberation ist keine rein Informationsverarbeitungsprozess; sie setzt auch ein Modus politischer Existenz: „Handeln“ voraus. Das im Sinne Arendts voraussetzt, dass Menschen im öffentlichen Raum erscheinen, miteinander sprechen und ihre Individualität zur Geltung bringen, sich gegenseitig anerkennen und Verantwortung für Worte und Handlungen übernehmen.

Wenn zentrale kommunikative Funktionen an KI delegiert werden, kann das als „Verschiebung“ von menschlichem verantwortlichem Handeln in Richtung einer funktionalisierten Prozesslogik verstanden werden, mit dem Risiko, dass Beteiligung zwar organisiert, aber die Erfahrung politischer Selbstwirksamkeit und Zurechenbarkeit geschwächt wird. Es ist wichtig, dass demokratisches Handeln nicht bloß moderiert werden soll, sondern sich als menschliches Handeln verantworten lassen. (Vgl. Waelen 2025, 4f)

Allerdings beziehen sich die bisher diskutierten Studien überwiegend auf die Wahrnehmung von KI im Allgemeinen und nicht spezifisch auf KI-Agenten. Neuere Entwicklungen im Bereich multi-agentischer Systeme betrachten KI-Agenten in deliberativen Kontexten nicht allein als unterstützend; Vielmehr bestätigen ihre funktional differenzierten Rollen innerhalb des Prozesses. Turkstra et al nennen als Beispiel, das System ArgsBase, in dem Verschiedene spezialisierte Agenten klar abgegrenzte Aufgaben erfüllen. Ein Moderator-Agent strukturiert

den Ablauf und organisiert das Turn-Taking, Deliberator-Agenten bringen Argumente ein und entwickeln diese weiter, während ein Analyzer-Agent den Diskurs auswertet, zusammenfasst und strukturiert. Der deliberative Prozess entsteht somit aus dem koordinierten Zusammenspiel mehrerer Agenten. (Vgl. Turkstra et al. 2026, 563-565)

Ein zentraler Vorteil dieses Ansatzes liegt in der Ausrichtung auf eine konstruktive Mensch-KI-Kollaboration, bei der KI nicht als Ersatz, sondern als Ergänzung menschlicher Urteilsbildung fungiert. Turkstra et al. stallen das in den Zusammenhang einer umfassenderen Konzeption „hybrider Intelligenz“, in der KI-Systeme darauf ausgelegt sind, menschliches Denken zu „erweitern“. Gleichzeitig knüpfen solche Systeme an Forderungen nach dialogbasierten Technologien an, die argumentatives Denken gezielt fördern sollen. In diesem Kontext wird auch auf das Konzept der „reasonable parrots“ von Musi et al. verwiesen, also Agenten, die sich an Diskussionen beteiligen und dabei grundlegende Prinzipien der Argumentation, wie Relevanz, Verantwortung und Freiheit, berücksichtigen. (Vgl. Turkstra et al. 2026, 563f)

Darüber hinaus werden weitere Vorteile hervorgehoben: Das System soll effektive Entscheidungsfindung ermöglichen, indem mehrere Sprachmodelle gemeinsam mit einem menschlichen Nutzer in einen strukturierten Dialog eingebunden werden und so unterschiedliche Stärken der einzelnen Modelle genutzt werden können. Zudem fördere das System menschliches kritisches Denken, unterstütze die Berücksichtigung multipler Perspektiven und erleichtere die Reflexion durch Echtzeit-Zusammenfassungen, offene Fragen und Argumentkarten, so Turkstra et al. Sie bezeichnen dies auch als relevant für öffentliche Beteiligungsprozesse sowie für die Forschung, etwa zur Analyse von Entscheidungsqualität oder kognitiven Prozessen; besonders hervorgehoben werden dabei die strukturierte Gesprächsführung und die Transparenz[25] des Diskussionsverlaufs. (Vgl. Turkstra et al. 2026, 563-568)

Turkstra et al. nennen auch mehrere Herausforderungen: Problematisch ist insbesondere die Stabilität längerer Interaktionen; mit zunehmender Dauer steigt die Komplexität, was zu inkonsistenten oder unerwarteten Antworten führen kann. Zudem zeigen die Modelle eine gewisse Verzerrungsanfälligkeit, etwa indem sie sich an Nutzerpositionen anpassen oder nur oberflächlich widersprechen, wodurch die angestrebte Vielfalt an Perspektiven eingeschränkt

25. Im Abschnitt 3.2.1. wurde bereits deutlich, dass mangelnde Transparenz eine zentrale Herausforderung beim Einsatz von KI-Agenten darstellt, insbesondere im Hinblick auf die Herkunft und Nutzung von Trainingsdaten.

werden kann. Auch die Gestaltung der Beiträge stellt eine Schwierigkeit dar: Zu kurze Antworten können die inhaltliche Qualität beeinträchtigen, während zu lange Beiträge den Gesprächsfluss stören. Darüber hinaus ist die praktische Wirksamkeit des Systems noch nicht ausreichend abgesichert und bedarf weiterer Evaluation. Schließlich kann die Komplexität des Systems für Nutzer als überfordernd empfunden werden, insbesondere da es auf strukturierte, reflektierte Deliberation und nicht auf schnelle Antworten ausgerichtet ist. (Vgl. Turkstra et al. 2026, 568f)

KI-Agenten lassen sich auch in deliberativen Kontexten nicht mehr lediglich als technische Mittel begreifen, da sie kommunikative Prozesse strukturieren, moderieren und sichtbar machen. Vielmehr übernehmen sie auch vermittelnde Funktionen innerhalb des Diskurses, wodurch menschliche Interaktionen strukturell geprägt und teilweise vorgeformt werden. In diesem Sinne erscheinen sie nicht mehr ausschließlich als Objekte politischer Praxis: Sie gestalten den deliberativen Prozess aktiv mit. Wird menschliches Handeln zunehmend durch technologische Systeme vermittelt, besteht die Gefahr, dass zentrale Bedingungen politischen Handelns unter Druck geraten. Deliberation droht damit, zwar formal organisiert zu werden, während Selbstwirksamkeit und gegenseitiger Anerkennung nicht in gleichem Maße erhalten bleiben.

3.4.2.3. Der Informationsraum unter Bedingungen generativer KI

Während Jungherr & Rauchfleisch vor allem die Beteiligungsseite in den Blick nehmen, analysieren Anand und Bentzen die Informationsraum-Seite demokratischer Stabilität: KI kann Transparenz und Zugang zu Information verbessern, gleichzeitig aber Desinformation, Manipulation und Überwachung verstärken. (Vgl. Anand 2025, 461-464; Bentzen 2025, 1-7)

Bentzen meint, dass generative KI selbst keine Täuschungsabsicht besitzt, aber menschlichen Akteuren ermöglicht, Einflusskampagnen überzeugender, schneller und schwerer erkennbar zu betreiben; entlang des gesamten Lebenszyklus (Daten, Prompts, Modellabstimmung, Output) bestehen Verzerrungs- und Manipulationspunkte. Deepfakes bedrohen demokratische Prozesse, indem sie Kandidaten diskreditieren, Wahlbetrug vortäuschen oder Narrative manipulieren; zugleich erodiert Vertrauen, weil Bürger Authentizität schlechter beurteilen können. (Vgl. Bentzen 2025, 1-5)

Besonders wichtig für diese Arbeit ist Bentzens Analyse von LLM-Chatbots, da sie in ihrer Funktionsweise agentische Eigenschaften aufweisen. Sie interagieren in Gesprächen und sind manipulierbar uns dadurch manipulativ, können falsche Inhalte erzeugen, reproduzieren. Zugleich zeigt Bentzen, dass Nutzer den Systemen menschliche Eigenschaften wie Verständnis, Humor oder moralisches Urteilen zuschreiben und ihnen teilweise sogar Empathie unterstellen, was das Vertrauen in ihre Ausgaben erhöht und ihre Wirkung im politischen Kontext verstärken kann. (Vgl. Bentzen 2025, 5f)

Hinzu kommen Risiken wie Informationswäsche (information laundering) und Datenvergiftung (data poisoning), durch die manipulierten Inhalte über scheinbar legitime Wissenskanäle in Trainingsdaten gelangen und von Chatbots reproduziert werden. Damit erscheinen KI-Agenten faktisch als skalierbare, automatisierte Interaktions- und Verbreitungsakteure im demokratischen Informationsraum, die Meinungsbildung, Vertrauen und Wahlprozesse beeinflussen und dadurch demokratische Prozesse gefährden können.

Im NED-Bericht von 2024 ist auch diese Diagnose zugespitzt, indem generative KI als ein Beschleuniger autoritärer Informationsmanipulation beschrieben wird: Sie reduziert Kosten, Zeit und technische Einstiegshürden für die Produktion und Verbreitung überzeugender synthetischer Inhalte in großem Maßstab und erleichtert damit langfristige Strategien, gesellschaftliches Vertrauen sowie die Idee einer geteilten, erkennbaren Wahrheit zu unterminieren. Zentral ist dabei der Mechanismus des „liar's dividend": Mit wachsender Verbreitung synthetischer Medien können kompromittierende Wahrheiten leichter als „fake" zurückgewiesen werden, was den Vertrauensverlust weiter verstärkt. Der Bericht zeigt zudem, dass autoritäre Akteure generative KI bereits in Wahlkontexten erproben und in ihre Strategien integrieren und dass solche Praktiken das Vertrauen in Online-Inhalte, Wahlen und demokratische Institutionen schwächen können. (Vgl. NED 2024)

Der Wahlzyklus 2024 wurde im Kontext rascher Fortschritte generativer KI als besonders herausfordernd für demokratische Prozesse beschrieben. Generative KI wirkt dabei vor allem als Verstärker bestehender Probleme wie Desinformation, Polarisierung und Vertrauensverlust und beschleunigt deren Verbreitung im digitalen Informationsraum. Daraus ergibt sich die Notwendigkeit verschiedener Gegenmaßnahmen, insbesondere in den Bereichen Transparenz, Regulierung, Plattformverantwortung und Förderung digitaler Kompetenzen. (Vgl. Farooq & de Vreese 2024)

In Verbindung mit Bentzen und dem NED-Bericht lässt sich zeigen, dass die Problematik nicht nur in einzelnen falschen Inhalten liegt; sie liegt eher in einer weitergehenden Erosion von Vertrauen und epistemischer Stabilität im demokratischen Informationsraum. Deepfakes und andere Formen synthetischer Medien beeinträchtigen die Qualität von Informationen und erschweren auch die Zuschreibung von Urheberschaft und Verantwortlichkeit, indem Authentizität zunehmend schwerer überprüfbar wird. (Vgl. Bentzen 2025, 2-5; NED 2024) In Anschluss an Waelen kann dies als Problem für zentrale Bedingungen politischen Handelns verstanden werden, da die Möglichkeit, Äußerungen eindeutig zuzuordnen und ihnen dauerhafte Bedeutung zuzuschreiben, unter Druck gerät. (Vgl. Waelen 2025, 4f) Der Mechanismus des „liar's dividend" verdeutlicht diese Dynamik: Wenn Wahrheits- und Falschheitszuschreibungen instabil werden, kann auch politische Rechenschaft untergraben werden (vgl. NED 2024).

Der wachsende Einsatz von KI-Agenten im demokratischen Informationsraum birgt die Gefahr einer schrittweisen Auslagerung menschlicher Urteilskraft. Insbesondere dort, wo ihre Entstehungsbedingungen und Funktionsweisen nicht ausreichend reflektiert werden, kann sich politische Orientierung zunehmend von eigenständigem Denken lösen. Problematisch wird dies vor allem dann, wenn KI-Systeme bestehende Machtinteressen verstärken und Aufmerksamkeit, Deutung und Sichtbarkeit gezielt strukturieren. Ein fehlender bewusster Umgang mit KI-generierten Inhalten begünstigt so Formen politischer Gedankenlosigkeit und kann langfristig zur Aushöhlung des öffentlichen Raums beitragen.

3.4.2.4. KI, Governance und demokratische Steuerungsfähigkeit

Der OECD[26]-Bericht zeigt, dass KI in staatlichen Kontexten unter anderem zur Verbesserung von Entscheidungsprozessen und zum besseren Verständnis gegenwärtiger Situationen eingesetzt wird, daneben auch für Prognosen sowie zur Optimierung von Informationsmanagement und Zugänglichkeit. Solche Anwendungen sollen politische Entscheidungsprozesse unterstützt haben und die Umsetzung von Politik sowie die Qualität öffentlicher Dienstleistungen verbessert haben. Als Beispiele hierfür wurden deliberative Tools wie Polis, die Konsensbereiche identifizieren, sowie KI-gestützte Assistenzsysteme im Bürgerservice genannt. (Vgl. OECD, 2025, 70)

26. Organisation for Economic Co-operation and Development

Der OECD-Bericht beschreibt frühe Formen sogenannter agentischer KI, die autonom operieren können. Dazu zählen unter anderem LLM-basierte Agenten, die eigenständig im Internet recherchieren und Informationen im Auftrag der Nutzenden interpretieren. Diese Systeme befinden sich jedoch noch in einem Entwicklungsstadium und weisen zahlreiche Einschränkungen sowie Risiken auf. Zugleich betonen viele Experten, dass Entscheidungen nicht an Maschinen delegiert werden sollten; vielmehr da solche Systeme faktisch Handlungsmacht im Informations- und Entscheidungsraum entfalten können, zielt die OECD-Analyse darauf ab, eine kooperative Verantwortungsordnung zwischen Menschen und Maschinen abzusichern. (Vgl. OECD 2025, 19-21)

Hinzu kommen Governance-Risiken wie Daten-Lock-in und Vendor-Lock-in. Restriktive Datenlizenzierungsvereinbarungen können dazu führen, dass relevante Daten nicht mit Entwicklern geteilt werden können, wodurch die Effektivität von KI-Systemen eingeschränkt wird. Gleichzeitig besteht die Gefahr eines Vendor-Lock-ins, bei dem staatliche Stellen von proprietären Technologien und Datenformaten einzelner Anbieter abhängig werden (vgl. OECD 2025, 210). Wenn staatliche Stellen zunehmend von agentischen KI-Systemen abhängig werden, die intransparent sind oder sich nur schwer kontrollieren lassen, geht ein Teil institutioneller Verlässlichkeit verloren. Politische Entscheidungen werden weniger nachvollziehbar und schlechter zurechenbar. Governance droht sich in Richtung technischer Sachzwänge zu verschieben, wodurch die demokratische Steuerungsfähigkeit faktisch eingeschränkt wird.

Die OECD hebt zugleich potenzielle Einsatzmöglichkeiten von KI im Bereich der öffentlichen Integrität hervor. KI kann dabei unterstützen, Muster in Daten zu erkennen, Beziehungen zwischen verschiedenen Akteuren zu analysieren sowie mithilfe agentenbasierter Modelle mögliche Auswirkungen von Antikorruptionsmaßnahmen vor deren Umsetzung zu simulieren. kann KI-gestützte Netzwerkanalyse dazu beitragen, Einflussstrukturen und potenzielle Interessenkonflikte sichtbar zu machen. Der Bericht betont jedoch, dass solche Anwendungen nur unter Bedingungen vertrauenswürdiger Nutzung eingesetzt werden sollten. Dazu gehören insbesondere Transparenz, Nachvollziehbarkeit und Erklärbarkeit der Entscheidungsprozesse sowie die Möglichkeit, Entscheidungen anzufechten. Zudem soll die letztliche Entscheidungsbefugnis auch bei zunehmender Autonomie von KI-Systemen beim Menschen verbleiben (vgl. 2025, 219f).

Diese Anforderungen sind jedoch nicht für alle menschlichen Akteure gleichermaßen einlösbar. Je komplexer und fortgeschrittener KI-Agenten werden, desto stärker verlagern sich

Verständnis, Kontrolle und effektive Anfechtbarkeit auf hochspezialisierte technische Expertise. Damit droht die demokratische Teilhabe in KI-integrierten Entscheidungsräumen implizit selektiv zu werden. Anstatt Inklusion zu fördern, kann sich die Kluft zwischen formaler Mitbestimmung und faktischer Einflussnahme vertiefen. So besteht die Gefahr, dass unter dem normativen Anspruch von Transparenz und Erklärbarkeit bestehende politische Machtstrukturen stabilisiert und technokratisch abgesichert werden.

Die Entwicklung leistungsfähiger KI-Systeme konzentriert sich derzeit auf eine sehr kleine Zahl globaler Unternehmen, insbesondere auf OpenAI, Google (inklusive DeepMind), Microsoft, Meta, Amazon (AWS) in den USA, sowie auf Baidu, Alibaba, Tencent und Nvidia in China. D. h. nur diese Firmen spielen eine international dominante Rolle im KI-Bereich. Insbesondere sogenannte KI-Agenten werden überwiegend von wenigen großen Technologieunternehmen in den USA und China entwickelt und kontrolliert. Diese Unternehmen sind privatwirtschaftlich organisiert und unterliegen keiner demokratischen Entscheidungsstruktur. Zentrale Entscheidungen über Trainingsdaten, Modellarchitekturen, Zugangsbedingungen oder Sicherheitsstandards werden innerhalb unternehmerischer Governance-Strukturen getroffen, nicht im Rahmen öffentlicher Deliberation. Wenn staatliche Institutionen KI-Agenten einsetzen, die auf diesen Modellen oder Infrastrukturen beruhen, entsteht faktisch ein Abhängigkeitsverhältnis. Demokratische Entscheidungsprozesse werden dann technisch vermittelt durch Systeme, deren grundlegende Bedingungen außerhalb demokratischer Kontrolle entstehen. Die scheinbar unvermeidliche Zunahme der Nutzung von KI-Agenten bedeutet daher nicht nur funktionale Unterstützung, sondern auch eine strukturelle Mitwirkung an bestehenden ökonomischen und machtpolitischen Konzentrationen im Technologiesektor.

3.4.2.5. Von KI-Vertrauen zu politischem Vertrauen und zurück

Fink-Hafner und Kaišić entwickeln ein konzeptionelles Modell, in dem Vertrauen in KI und politisches Vertrauen in einem wechselseitigen, teilweise indirekten Verhältnis zueinanderstehen. Vertrauen in KI wird dabei als kontingent und veränderlich beschrieben und durch ein Bündel von Faktoren geprägt. Dazu zählen demografische Merkmale,

nutzerbezogene Faktoren,[27] technologiebezogene Eigenschaften[28] sowie kontextuelle Einflüsse, darunter soziale Normen, Medienrepräsentationen von KI und institutionelle Sicherungsmechanismen. (Vgl. Fink-Hafner und Kaišić 2025: 353-358)

Dabei wird politisches Vertrauen als relationales Konzept verstanden, das die Beziehung zwischen Bürgern, politischen Institutionen und dem demokratischen System beschreibt. Es wird durch ein Zusammenspiel von Mikro- und Makrofaktoren beeinflusst. Auf der Mikroebene sind individuelle Merkmale und sozioökonomische Erfahrungen relevant, etwa Bildungsniveau, Arbeitsmarktposition oder wirtschaftliche Sicherheit. Zentral ist dabei, dass politisches Vertrauen vor allem auf Wahrnehmungen und Bewertungen beruht und nicht zwingend auf objektiv messbaren Leistungsindikatoren staatlichen Handelns.
Auf der Makroebene wirken strukturelle und kontextuelle Faktoren wie Korruption, sozioökonomische Ungleichheiten, institutionelle Inklusivität, die Qualität öffentlicher Dienstleistungen sowie demokratische Stabilität und Rechtsstaatlichkeit. Auch Medien spielen eine wichtige Rolle, da ihre Berichterstattung und Rahmung politischen Handelns beeinflussen, wie politische Akteure, Institutionen und Entscheidungsprozesse wahrgenommen werden. (Vgl. Fink-Hafner und Kaišić 2025: 359 f.)

Im Modell wird gezeigt, dass Misstrauen gegenüber KI politisches Vertrauen untergraben kann, etwa wenn KI als Faktor wahrgenommen wird, der Rechenschaftspflichten erschwert, politische Repräsentation verzerrt, Vertrauen in Medien schwächt oder politische Beteiligung kompliziert. Umgekehrt kann hohes politisches Vertrauen die Akzeptanz staatlicher KI-Anwendungen fördern, da Bürger der Regierung eher zutrauen, KI verantwortungsvoll einzusetzen und zu regulieren (vgl. Fink-Hafner & Kaišić 2025, 360-362).

Wir kommen zu dem Ergebnis, dass die Legitimität von KI-Agenten nicht vorausgesetzt werden kann, wenn sie in demokratisch besonders sensiblen Bereichen eingesetzt werden. Öffentliche Akzeptanz muss vielmehr durch eine verantwortungsvolle Gestaltung agentischer Systeme aktiv hergestellt werden, insbesondere durch Transparenz, Nachvollziehbarkeit, Anfechtbarkeit und klar zugewiesene menschliche Verantwortung, nicht durch kommunikative oder symbolische Vertrauensstrategien.

27. etwa Erfahrungen, Einstellungen, Werte und wahrgenommene Kontrolle

28. wie Robustheit, Transparenz, Erklärbarkeit, Fairness, Datenschutz und Rechenschaftspflicht

Dabei geht es nicht bloß um Vertrauen in die technische Korrektheit; es geht auch um die Stabilität eines gemeinsamen Raums, in dem Aussagen überprüfbar bleiben, Verantwortlichkeiten eindeutig zugeschrieben werden können und soziale Bezogenheit gewährleistet ist. Der Einsatz von KI-Agenten greift unmittelbar in diese Bedingungen politischer Öffentlichkeit ein und strukturiert sie mit. Ein reflektierter und verantwortungsvoller Umgang mit solchen Systemen ist daher zentral, stößt jedoch an Grenzen: Mit zunehmender Komplexität und Autonomie von KI-Agenten sinkt die Nachvollziehbarkeit ihrer Entscheidungsprozesse. Vertrauen wird unter diesen Bedingungen prekär und kann als strukturelle Voraussetzung politischer Öffentlichkeit selbst geschwächt werden.

3.4.2.6. Rechtsrahmen und demokratische Legitimität

Bentzen stellt heraus, dass die EU auf KI-gestützte Informationsmanipulation mit einem menschenzentrierten Ansatz und einer grundrechtsbasierten digitalen Regulierung reagiert. Zu den relevanten Rechtsinstrumenten zählen insbesondere der AI Act, der DSA einschließlich des Code of Conduct on Disinformation, die Richtlinie zur Bekämpfung von Gewalt gegen Frauen sowie der EMFA. Zugleich betont Bentzen, dass die Transparenzpflichten des AI Act nicht genügen, um manipulative Wirkungen auszuschließen, da auch gekennzeichnete Deepfakes oder KI-generierte Inhalte individuelle Autonomie, informierte Entscheidungsfindung und demokratische Prozesse beeinträchtigen können. (Vgl. Bentzen 2025, 9-12)

Die Europäische Kommission führt „AI solutions used in the administration of justice and democratic processes“ ausdrücklich als Hochrisiko-Anwendungsbereich an (vgl. Europäische Kommission 2026). Die FRA weist in diesem Zusammenhang auf einen erheblichen Umsetzungsbedarf hin: Zahlreiche Anbieter und Anwender erfassen Grundrechtsrisiken bislang nicht systematisch; zudem besteht die Gefahr regulatorischer Schlupflöcher infolge weiter Begriffsbestimmungen sowie einer starken Abstützung auf Selbstbewertungen. Gefordert werden daher praxisnahe Leitlinien, verbindliche Grundrechte-Folgenabschätzungen, externe Kontrollmechanismen sowie eine angemessene Ausstattung der zuständigen Aufsichtsbehörden. (Vgl. FRA 2025, 5; 8-13)

Der bestehende Rechtsrahmen reagiert damit implizit auf politisch bedeutsame Gefährdungslagen: Wenn der öffentliche Raum technisch so strukturiert wird, dass Wahrheit,

Autorenschaft und Verantwortungszuschreibung unsicher werden, berührt dies die Grundlagen demokratischer Freiheit.

Der reine Informationszugang ist jedoch nicht notwendigerweise demokratiefördernd. Demokratie verlangt mehr als die bloße Verfügbarkeit von Daten; sie setzt faire Diskursbedingungen, Transparenz, Nachvollziehbarkeit, gleiche Partizipationsmöglichkeiten sowie geteilte Verantwortlichkeit voraus. KI-Agenten können zwar die Verbreitung von Informationen beschleunigen, zugleich jedoch den öffentlichen Diskurs beeinflussen oder verzerren. Eine solche Personalisierung kann der für demokratische Ordnungen konstitutiven Pluralität zuwiderlaufen und dadurch die freie und gleiche Willensbildung beeinträchtigen. Beruhen Entscheidungen zudem auf intransparenten, sogenannten Black-Box-Systemen, entsteht ein Spannungsverhältnis zur demokratischen Forderung nach öffentlicher Begründung und kontrollierbarer Entscheidungsfindung.

Davon abgesehen bestehen Risiken struktureller Diskriminierung infolge verzerrter Trainingsdaten sowie Probleme ungleicher Zugangsbedingungen zu digitalen Technologien: Unterschiedliche digitale Kompetenzen und technische Ressourcen können bestehende soziale Ungleichheiten verstärken. Hinzu kommt das Problem automation bias[29] ins Spiel. Auch wenn formal menschliche Aufsicht vorgesehen ist, neigen Nutzer dazu, maschinellen Ausgaben besonderes Gewicht beizumessen. Ist diese menschliche Kontrolle jedoch nur eingeschränkt wirksam oder ohne reale Eingriffsmöglichkeiten ausgestaltet, droht eine Diffusion von Verantwortung. Bei KI-Agenten verschärft sich diese Problematik: Sie treten autonom im gesellschaftlichen Raum auf und erscheinen als handlungsähnliche Akteure, ohne selbst Träger rechtlicher oder moralischer Verantwortung zu sein. Bleiben die Verantwortungs- und Zurechnungsstrukturen zwischen Entwicklern, Betreibern und Anwendern unklar, entsteht ein demokratietheoretisches Defizit, das die Legitimität solcher Systeme grundlegend infrage stellt.

3.4.2.7. KI-Agenten als politische Mitglieder – unter kooperativer Verantwortung?

Die Fähigkeit von KI-Agenten, als autonom agierende Systeme im öffentlichen Raum aufzutreten, wirft die grundlegende Frage nach ihrer demokratietheoretischen Verortung auf. Indem sie kommunikative Prozesse strukturieren, Informationen selektieren oder

29. Das Problem eines Übervertrauens in algorithmisch generierte Ergebnisse

Entscheidungsprozesse vorbereiten, entfalten sie eine faktische Handlungsmacht innerhalb demokratischer Ordnungen. Daraus ergibt sich die Frage, ob und in welcher Weise sie als aktive Mitakteure demokratischer Praxis zu begreifen sind.

Durch ihre Funktionalität tragen solche Systeme zur weiteren Dominanz instrumenteller und konsumorientierter Logiken bei, wodurch Arbeit und insbesondere Handeln im Sinne Arendts zurückgedrängt werden. Das droht Sinnverlust, Entpolitisierung und, vermittelt über Konsumlogiken, eine weitere Erosion gemeinsamer Weltbezüge. (Vgl. Waelen 2025, 8-12)

Das macht deutlich, dass KI-Agenten nicht lediglich als passive Instrumente technischer Steuerung verstanden werden können. Ich denke, es ist problematisch, sie als bloße „Opferobjekte" technischer Steuerung zu betrachten, während sie faktisch politische Kommunikations- und Entscheidungsprozesse mitprägen. Entscheidend ist daher die Entwicklung eines Rahmens, der ihre funktionale Autonomie anerkennt, zugleich jedoch klare Verantwortungs- und Kontrollmechanismen etabliert, um die Integrität der demokratischen Ordnung zu sichern.

Sinnvoll erscheint ein Modell, das KI nicht als Subjekte im klassischen Sinne versteht (vgl. Jungherr & Rauchfleisch 2025, 2f), stattdessen als wirkmächtige Elemente innerhalb einer Mensch-Maschine-Konstellation. Eine solche Konzeption müsste klären, wie agentische Systeme in demokratische Ordnungen integriert werden können, ohne ihnen personale Eigenschaften wie Bewusstsein oder moralische Selbstgesetzgebung zuzuschreiben. Politische Mitgliedschaft im strengen Sinne erschöpft sich nicht in bloßer Funktionalität; sie setzt vielmehr die Teilnahme an einer Praxis von Öffentlichkeit, Pluralität und Verantwortlichkeit voraus: Sie betrifft heutige KI-Agenten nicht.

An dieser Stelle ist eine Erinnerung an die Technikphilosophie von Martin Heidegger bedeutungsvoll: Wie besprochen,[30] Heidegger beschreibt die neuzeitliche Technik nicht bloß als Instrument, sondern als eine spezifische Weise des Entbergens, die er als „Gestell" bezeichnet. Im Modus des Gestells erscheint die Welt nicht mehr als etwas, das sich eigenständig zeigt, sondern als etwas, das als „Bestand" verfügbar gemacht wird. Natur wird zum „Bestellbaren"; sie wird herausgefordert, berechnet, gespeichert und abrufbar gehalten. Das Seiende tritt primär unter dem Gesichtspunkt seiner Nutzbarkeit und Disponibilität in Erscheinung.

30. Im Teil 2.2.3. Prozessuale Emergenz der KI: Zwischen dynamis und energeia

KI-Agenten intensivieren diese Logik des Bestellbaren erheblich. Sie machen materielle Ressourcen, Informationen, Präferenzen, Kommunikationsprozesse und Entscheidungsstrukturen jederzeit abrufbar und bestellbar. Die Welt wird durch sie in gesteigertem Maße verfügbar und bestellbar. So werden wie Naturressourcen, auch soziale Beziehungen, Aufmerksamkeit, Meinungen und Verhaltensmuster in einen Modus der Berechenbarkeit überführt. KI-Agenten täuschen Menschen, indem sie zunehmend den Eindruck erwecken, bloße Objekte oder Werkzeuge zu sein. In der Tat tragen sie aufgrund ihres erheblichen Beitrags zur allgemeinen Bestellbarkeit faktisch zu einer Entfremdung des Menschen von der Welt und von sich selbst bei.

KI-Agenten sind dazu da, mit uns Arm in Arm die oberflächliche Betrachtungsweise der Welt durchzusetzen. Sie tragen dazu bei, dass Wirklichkeit zunehmend als bestellbar erscheint. Sie strukturieren Informationsflüsse, priorisieren Inhalte, automatisieren Entscheidungen und erzeugen damit eine Ordnung, in der Handlungsoptionen vorselektiert sind. Ein Problem ist, dass diese Vorstrukturierung nie neutral ist. Diese Agenten sind keine Subjekte im starken Sinne, wohl aber operative Knotenpunkte in einem Geflecht, das öffentliche Wirklichkeit hervorbringt.

Wenn KI-Agenten als politische „Mitglieder“ bezeichnet werden, folgt daraus nicht, dass ihnen ein grundrechtlicher Status zukommt. Vielmehr bleibt die Zuerkennung von Grundrechten, wie auch im Kontext des EMFA betont wird an menschliche Subjekte gebunden (vgl. Bentzen 2025, 11f). KI-Agenten sind daher keine Rechtsträger, aber funktional wirksame Akteure innerhalb demokratischer Verfahren zu verstehen. Ihre Autonomie ist technisch und zweckgebunden; deshalb kann Verantwortung nicht bloß bei ihnen selbst liegen. Die Rede von einer kooperativen Verantwortung ist sinnvoll, unter der KI-Agenten demokratisch integrierbar werden können. KI darf moderieren, strukturieren und unterstützen, aber eine klare menschliche Zuständigkeit muss eindeutig verankert bleiben (Vgl. OECD 2025, 219f; FRA 2025, 57).

Politische Mitgliedschaft von KI-Agenten ist somit keine Frage symbolischer Anerkennung; dabei geht es um die Frage regulierter Rolle. Wo ihre Wirksamkeit transparent gekennzeic net, erklärbar, überprüfbar und institutionell verantwortet sowie inklusiv zugänglich und diskriminierungsarm gestaltet ist, können sie demokratische Verfahren stärken. Wo dies fehlt, drohen Manipulationsskalierung, Vertrauensverlust und die Verwandlung politischer Praxis in eine bloß funktionale Dienstleistung.

4. Fazit und Ausblick

Die vorliegende Arbeit hat gezeigt, dass KI-Agenten nicht als bloße Weiterentwicklung technischer Objekte verstanden werden können; sie stellen eine eigenständige, hybride Form von Akteurialität dar. Ihre ontologische Zwischenstellung, zwischen Künstlichkeit und Natürlichkeit, zwischen menschlicher Konstruktion und systemischer Eigenständigkeit, macht sie zu relationalen Grenzphänomenen, die sich klassischen Kategorien entziehen.

Ihre besondere Seinsweise bildet die Grundlage für die komplexen ethischen Herausforderungen, die mit ihrem Einsatz verbunden sind. Anders als klassische KI-Systeme verfügen KI-Agenten über ein erhöhtes Maß an Autonomie, strategischer Handlungsfähigkeit und sozialer Einbettung, wodurch Fragen nach Verantwortung, Vertrauen, Kontrolle und menschlicher Autonomie in neuer Schärfe auftreten.

Die Analyse hat deutlich gemacht, dass sich diese Herausforderungen nicht auf eine einzelne Dimension reduzieren lassen, vielmehr technische, philosophische sowie sozio-politische Aspekte gleichermaßen betreffen. Insbesondere der Umgang mit Daten, die Gefahr von Verantwortungslücken, anthropomorphe Fehlzuschreibungen sowie Auswirkungen auf demokratische Prozesse zeigen, dass KI-Agenten tief in gesellschaftliche Strukturen eingreifen und diese aktiv mitgestalten.

Daraus folgt, dass eine rein technologische Perspektive unzureichend ist. Vielmehr bedarf es einer interdisziplinären Reflexion, die sowohl die ontologischen und ethischen Grundlagen als auch die praktischen Konsequenzen berücksichtigt. Nur durch eine bewusste und kritische Auseinandersetzung kann das Potenzial von KI-Agenten verantwortungsvoll genutzt und gleichzeitig ihren Risiken angemessen begegnet werden.

KI-Agenten tragen zur Neubestimmung zentraler philosophischer Begriffe wie Handlung, Verantwortung und Autonomie bei. Die zukünftige Einbettung von KI-Agenten in gesellschaftliche Strukturen hängt maßgeblich davon ab, inwiefern es gelingt, ihre Gestaltung normativ zu rahmen und sie in eine ethisch tragfähige gesellschaftliche Ordnung zu integrieren.

5. Literaturverzeichnis

Adrià Puigdomènech, Bilal; Piot, Steven; Kapturowski, Pablo; Sprechmann, Alex; Vitvitskyi, Daniel; Guo, Charles; Blundell, Charles: „Agent57: Outperforming the Human Atari Benchmark“. In: *Google DeepMind*. 2020. https://deepmind.google/discover/blog/agent57-outperforming-the-human-atari-benchmark/ (Letzter Zugriff: 28.06.2025).

Amirizaniani, Maryam: „Enhancing AI Agents with Human Theory of Mind (ToM) for Context-Aware Actions“. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR ’25). Padua 2025. https://doi.org/10.1145/3726302.3730127 (Zugriff: 02.01.2026).

Anand, P.: „AI: Challenges for Democracy and Some Policy Solutions“. In: *Critical Review of International Social and Political Philosophy.* 2025. https://www.tandfonline.com/doi/full/10.1080/19452829.2025.2518307 (Letzter Zugriff: 01.02.2026).

Arendt, Hannah: „Kultur und Politik“. In: *Dies.: Zwischen Vergangenheit und Zukunft*. Hrsg. von Ursula Ludz. München 1994.

Arendt, Hannah: *Vita activa oder Vom tätigen Leben*. München: Piper 2008.

Arendt, Hannah: *Was ist Politik? Fragmente aus dem Nachlass*. München: Piper 2007.

Aristoteles: *De Caelo;* Über den Himmel. Übersetzt und kommentiert von H. D. P. Lee. Loeb Classical Library 1952.

Aristoteles: *Metaphysik*. Übersetzt und hrsg. von Franz Dirlmeier. 7. Auflage. Hamburg: Felix Meiner Verlag 1995.

Aristoteles: *Physik*. Übersetzt und hrsg. von Hans Günter Zekl. Hamburg: Felix Meiner Verlag 1987.

Averly, R.; Chao, W.-L.: „Unified Out-Of-Distribution Detection: A Model-Specific Perspective“. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV*). 2023, 1453–1463. https://openaccess.thecvf.com/content/ICCV2023/html/Averly_Unified_Out-Of-Distribution_Detection_A_Model-Specific_Perspective_ICCV_2023_paper.html (Letzter Zugriff: 16.03.2026).

Bentzen, Naja: „Information Manipulation in the Age of Generative Artificial Intelligence“. In: *European Parliamentary Research Service (EPRS), Members’ Research Service*, PE 779.259. Dezember 2025. https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/779259/EPRS_BRI%282025%29779259_EN.pdf (Letzter Zugriff: 01.02.2026).

Birnbacher, Dieter: „Natürlich-künstlich“. In: Neuhäuser, Christian; Raters, Marie-Luise; Stoecker, Ralf (Hg.): *Handbuch Angewandte Ethik*. 2. Auflage. Germany: J. B. Metzler’sche Verlagsbuchhandlung & Carl Ernst Poeschel GmbH 2023.

Birnbacher, Dieter: *Natürlichkeit*. Berlin: Walter de Gruyter GmbH & Co. KG 2006.

Bl2run: „KI-Assistenten vs. KI-Agenten: Was sind die Unterschiede?“. 2025. https://bi2run.de/blog/ki-assistenten-vs-ki-agenten-was-sind-die-unterschiede/ (Letzter Zugriff: 14.07.2025).

Blum, André L.: „Der intelligente Körper und sein Hirn“. In: Blum, André L.; Krois, John Michael; Rheinberger, Hans-Jörg (Hg.): *Verkörperungen*. Berlin/Boston: De Gruyter 2012, 11–46.

Bradley, Adam: *Artificial Language Agents, Representation, and the Limits of Mental State Attribution.* Manuskript 2025.

Calvo, Rafael A.; Peters, Dorian; Vold, Karina; Ryan, Richard M.: „Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry“. In: Burr, Christopher; Floridi, Luciano (Hg.): *Ethics of Digital Well-Being. Philosophical Studies Series 140*. Cham: Springer 2020, 31–54. https://doi.org/10.1007/978-3-030-50585-1_2 (Letzter Zugriff: 22.01.2026).

Cappelen, Herman; Dever, Josh: „Going Whole Hog: A Philosophical Defense of AI Cognition“. Manuskript / Preprint 2024/2025.

Christen, M.: „Autonomie eine Aufgabe für die Philosophie“. In: *Studia Philosophica* 66, 2007, 175–196.

Cody, Steven: „OpenAI, Altman Sued Over ChatGPT’s Role in California Teen’s Suicide“. In: *Reuters*, 26.08.2025. https://www.reuters.com/sustainability/boards-policy-regulation/openai-altman-sued-over-chatgpts-role-california-teens-suicide-2025-08-26/ (Letzter Zugriff: 28.08.2025).

Council of Economic Advisers: Artificial Intelligence and the Great Divergence. Washington, D.C. 2026. https://www.whitehouse.gov/research/2026/01/artificial-intelligence-and-the-great-divergence/ (Letzter Zugriff: 28.01.2026).

Crawford, Kate: *Atlas of AI. Power, politics, and the planetary costs of artificial intelligence*. New Haven, London: Yale University Press 2021.

De Ridder, Alexander: „Types of Agent Communication Languages“. 2025. https://smythos.com/ai-agents/ai-agent-development/types-of-agent-communication-languages/ (Letzter Zugriff: 25.05.2025).

Diels, Hermann; Kranz, Walther (Hrsg.): *Die Fragmente der Vorsokratiker. Griechisch und Deutsch*. 6. Auflage. Berlin: Weidmann 1951–1952; ND Dublin/Zürich: Weidmann 1989.

Elsakr, Hannah: „Adobe Firefly und Adobe Express verfügen nun über das Flash Image Model Gemini 2.5 von Google“. In: *Adobe Blog News*, 26.08.2025. https://blog.adobe.com/de/publish/2025/08/26/adobe-firefly-und-adobe-express-verfuegen-nun-ueber-das-flash-image-model-gemini-2-5-von-google (Letzter Zugriff: 28.08.2025).

European Commission: „AI Act“. Brüssel, laufend aktualisiert. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (Letzter Zugriff: 05.02.2026).

European Commission: *AI Act Explorer. Article 5: Prohibited AI Practices. Artificial Intelligence Act* (Regulation (EU) 2024/1689), official version of 13 June 2024. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-5 (Letzter Zugriff: 13.12.2025).

European Union Agency for Fundamental Rights (FRA): *Assessing High-risk Artificial Intelligence: Fundamental Rights Implications.* Wien 2025. https://fra.europa.eu/en/publication/2025/assessing-high-risk-ai (Letzter Zugriff: 05.02.2026).

Farooq, Aqsa; Claes de Vreese; Co-chairs of the EDMO Internal Expert Group on Generative AI and Disinformation: „Generative AI and Disinformation in 2024 Elections: Implications for Democracy Going Forward“. Bericht, Dezember 2024. https://edmo.eu/blog/generative-ai-and-disinformation-in-2024-elections-implications-for-democracy-going-forward/ (Letzter Zugriff: 21.02.2026).

Fink-Hafner, Katarina Kaisic: „Artificial Intelligence and Political Trust“. In: *Politics in Central Europe* 21 (3), 2025. https://www.politicsincentraleurope.eu/images/_PCE/2025/PCE_21-3/Fink-Hafner_et_al_PCE_21_2025_3.pdf (Letzter Zugriff: 01.02.2026).

Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; Schäfer, B.; Valcke, P.; Vayena, E.: „AI4People’s Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.“ *White Paper*. Brussels: Atomium – European

Institute for Science, Media and Democracy (EISMD) 2018. https://ai4people.org/PDF/AI4People_Ethical_Framework_For_A_Good_AI_Society.pdf (Letzter Zugriff: 16.03.2026).

Gabriel, Iason; Manzini, Arianna; Keeling, Geoff; Hendricks, Lisa Anne; Rieser, Verena; Iqbal, Hasan et al. *The Ethics of Advanced AI Assistants.* Google DeepMind 2024. https://doi.org/10.48550/arXiv.2404.16244.

Gesnot, Rénald: *The Impact of Artificial Intelligence on Human Thought*. Manuskript 2025.

Gloy, Karen: „Die drei Grundsätze aus Fichtes; Grundlage der gesamten Wissenschaftslehre‘ von 1794“. In: *Philosophisches Jahrbuch* 126, 2019, 289–307.

Goldman Sachs: „What to Expect From AI in 2026: Personal Agents, Mega Alliances, and the Gigawatt Ceiling“. In: *Goldman Sachs Insights/Outlooks*, 22. Jänner 2026. https://www.goldmansachs.com/insights/articles/what-to-expect-from-ai-in-2026-personal-agents-mega-alliances (Letzter Zugriff: 30.01.2026).

Google: „Gemini Thinking | Gemini API | Google AI for Developers“. 2025. https://ai.google.dev/gemini-api/docs/thinking (Letzter Zugriff: 20.08.2025).

Gramelsberger, Gabriele: *Philosophie des Digitalen zur Einführung*. Hamburg: Junius Verlag GmbH 2023.

Green, Brian Patrick: „Artificial Intelligence and Ethics: Sixteen Challenges and Opportunities“. In: Santa Clara Markkula Center for Applied Ethics. https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics-sixteen-challenges-and-opportunities/ (Letzter Zugriff: 22.06.2025).

Hadfield, Gillian K.; Koh, Andrew: „An Economy of AI Agents“. In: arXiv preprint, arXiv:2509.01063 [econ.GN]. 2025. https://arxiv.org/abs/2509.01063 (Letzter Zugriff: 20.01.2026).

Hawkins, Jeff: „What Intelligent Machines Need to Learn From the Neocortex“. In: *Institute of Electrical and Electronics Engineers*, Inc. 2017.

Hawley, Katherine: „Trust, Distrust and Commitment“. In: *Noûs* 48 (1), 2014, 1–20. doi: 10.1111/nous.12000. URL: https://onlinelibrary.wiley.com/doi/10.1111/nous.12000 (Letzter Zugriff: 02.01.2025).

Heidegger, Martin: „Die Frage nach der Technik“. In: *Gesamtausgabe. I. Abteilung: Veröffentlichte Schriften 1910–1976. Band 7: Vorträge und Aufsätze*. Frankfurt am Main: Vittorio Klostermann GmbH 2000, 5–37.

Heidegger, Martin: Sein und Zeit. 19. Auflage. Tübingen: Max Niemeyer Verlag 2006.

Hu, Charlotte; Downie, Amanda; Finio, Matthew: „AI Agents vs. AI Assistants“. In*: IBM*. 2025. https://www.ibm.com/think/topics/ai-agents-vs-ai-assistants (Letzter Zugriff: 14.07.2025).

Hu, Charlotte; Downie, Amanda; Finio, Matthew. „AI Agents vs. AI Assistants“. *IBM Think*. 2025. https://www.ibm.com/think/topics/ai-agents-vs-ai-assistants. (Letzter Zugriff: 27.05.2025)

Imhof, Margarete: *Psychologie für Lehramtsstudierende.* 2. Auflage. Wiesbaden: VS Verlag für Sozialwissenschaften 2011.

Jones, Charles I.: „A.I. and Our Economic Future“. In: *NBER Working Paper* 34779. National Bureau of Economic Research 2026. https://www.nber.org/papers/w34779 (Letzter Zugriff: 20.03.2026).

Jungherr, Andreas: „Artificial Intelligence in Deliberation: The AI Penalty and Democratic Legitimacy“. In: *Journal of Communication*. 2025. https://www.sciencedirect.com/science/article/pii/S0740624X25000735 (Letzter Zugriff: 01.02.2026).

Kandikatla, L.; Radeljić, B.: „AI and Human Oversight: A Risk-Based Framework for Alignment“. In: arXiv preprint, arXiv:2510.09090 [cs.AI]. 2025. https://arxiv.org/abs/2510.09090 (Letzter Zugriff: 23.01.2026).

Kandikatla, Laxmiraju; Radeljić, Branislav: „AI and Human Oversight: A Risk-Based Framework for Alignment“. *Working Paper*. In: arXiv Preprint. 2025. https://arxiv.org/pdf/2510.09090 (Letzter Zugriff: 20.01.2026).

Kant, Immanuel: *Kritik der Urteilskraft.* In: *Kants Werke. Akademie-Textausgabe.* Bd. 5. Berlin/New York 1968, 165–486.

Kergroach, Sandrine; Héritier, Julien: „Early Divides in the Transition to Artificial Intelligence“. In: *OECD Regional Development Papers,* Nr. 147. Paris: OECD Publishing 2025. https://doi.org/10.1787/7376c776-en (Letzter Zugriff: 10.02.2026).

KI-Navi: „KI Assistent vs. KI Agent; kennst du den entscheidenden Unterschied“. 10.07.2025. https://www.ki-navi.com/2025/07/08/ki-assistent-vs-ki-agent-erklaerung-zum-entscheidenden-unterschied/?utm_source=chatgpt.com (Letzter Zugriff: 27.09.2025).

Kim, Kyung-Hoon; Sugiyama, Satoshi: „AI Romance Blooms as Japan Woman Weds Virtual Partner of Her Dreams“. In: *Reuters*. 16.12.2025 (Letzter Zugriff: 23.12.2025).

Kim, Myung Ho: *Executable Epistemology: The Structured Cognitive Loop as an Architecture of Intentional Understanding*. Manuskript 2025.

Koch, Christof: „Christof Koch über Bewusstsein Phi II Teil 1 von 3. Persönliches Interview“. Geführt von Arvid Leyh. Veröffentlicht am 02.10.2012. https://www.youtube.com/watch?v=6bSAe3ykw_o (Letzter Zugriff: 07.04.2025).

Koch, Christof: *Bewusstsein- Warum es weit verbreitet ist, aber nicht digitalisiert werden kann*. Berlin/ Heidelberg: Springer 2020.

Koralus, P.: „The Philosophic Turn for AI Agents: Replacing Centralised Digital Rhetoric with Decentralised Truth-Seeking“. In: arXiv preprint, arXiv:2504.18601 [cs.AI]. 2025. https://arxiv.org/abs/2504.18601 (Letzter Zugriff: 22.01.2026).

Koralus, Philipp: „The Philosophic Turn for AI Agents: Replacing Centralized Digital Rhetoric with Decentralized Truth-Seeking“. Manuskript 2025.

Korinek, Anton: „AI Agents for Economic Research“. In: *NBER Working Paper* 34202. Cambridge, MA: National Bureau of Economic Research 2025. http://www.nber.org/papers/w34202 (Letzter Zugriff: 28.01.2026).

Larghi, Silvia; Datteri, Edoardo: „Mentalistic Stances towards AI Systems“. Manuskript / Preprint 2024.

League of Conservation Voters: „Decoding Trump’s Energy Policy: Misleading Terms and What They Really Mean“. In: *League of Conservation Voters*, 13.02.2025. https://www.lcv.org/blog/decoding-trumps-energy-policy-misleading-terms-and-what-they-really-mean/ (Letzter Zugriff: 01.03.2025).

MakeMyTrip: „MakeMyTrip Launches GenAI-Trip-Planning Assistant Myra“. In: *Moneycontrol*, 07.09.2025. https://www.moneycontrol.com/technology/makemytrip-rolls-out-new-multilingual-genai-assistant-article-13412116.html (Letzter Zugriff: 28.09.2025).

Metzinger, Thomas (Hrsg.): *Grundkurs Philosophie des Geistes. Band 3: Intentionalität und mentale Repräsentation.* Paderborn: mentis Verlag GmbH 2010.

Moltbook: „Social Network für KI-Agenten“. https://www.moltbook.com/post/35e3ca66-f1bd-4517-a2a6-a3f2e950eb2c (Letzter Zugriff: 11.03.2026).

Müller, Vincent C.: „Philosophy of AI: A Structured Overview“. In: Smuha, Nathalie A. (Hrg.): *Cambridge Handbook on the Law, Ethics and Policy of Artificial Intelligence.* Cambridge: Cambridge University Press 2025. https://doi.org/10.1017/9781009367783.004 (Zugriff: 02.03.2026).

NED (National Endowment for Democracy): „Manufacturing Deceit: How Generative AI Supercharges Information Manipulation“. Bericht von Beatriz Saab. *International Forum for Democratic Studies*. Washington, DC 2024. https://www.ned.org/manufacturing-deceit-how-generative-ai-supercharges-information-manipulation/ (Letzter Zugriff: 21.02.2026).

NIST (National Institute of Standards and Technology): *Managing Misuse Risk for Dual-Use Foundation Models*. NIST AI 800-1, Second Public Draft. U.S. AI Safety Institute, U.S. Department of Commerce 2025. https://doi.org/10.6028/NIST.AI.800-1.2pd (Letzter Zugriff: 09.02.2026).

Nyholm, Sven: „Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility Loci“. In: *Science and Engineering Ethics* 24 (4), 2018, 1201–1219. https://doi.org/10.1007/s11948-017-9943-x (Letzter Zugriff: 10.11.2025).

Nyholm, Sven: „Responsibility Gaps, Value Alignment, and Meaningful Human Control over Artificial Intelligence“. In: Placani, Adriana; Broadhead, Stearns (Hrsg.): *Risk and Responsibility in Context*. 1. Auflage. Routledge 2023, 191–213.

Nyrup, R.: „Trustworthy AI: A Plea for Modest Anthropocentrism“. In: *Asian Journal of Philosophy* 2 (40), 2023. https://doi.org/10.1007/s44204-023-00096-w (Letzter Zugriff: 01.01.2026).

O’Gieblyn, Meghan: „Is the Internet Conscious? If It Were, How Would We Know?“. Veröffentlicht am 16.09.2020. https://www.wired.com/story/is-the-internet-conscious-if-it-were-how-would-we-know/ (Letzter Zugriff: 21.03.2025).

OECD: *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*. Paris: OECD Publishing 2025. https://doi.org/10.1787/795de142-en (Letzter Zugriff: 05.02.2026).

Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; Bernstein, M. S.: „Generative Agents: Interactive Simulacra of Human Behavior“. 2023. https://arxiv.org/abs/2304.03442 (Letzter Zugriff: 11.03.2026).

Pasetti, M.; Santos, J. W.; Kluge Corrêa, N.; de Oliveira, N.; Palhares Barbosa, C.: „Technical, Legal, and Ethical Challenges of Generative Artificial Intelligence: An Analysis of the Governance of Training Data and Copyrights“. In: *RAIES – Rede de Inteligência Artificial Ética e Segura*. 2025.

Roshan, Gya; Li, Gathy: *AI Agents in Action: Foundations for Evaluation and Governance*. White Paper. Geneva: World Economic Forum 2025.

Schurr, Philipp: „Künstliche Intelligenz“. In: *mind square; IT Beratung und Entwicklung*. 05.03.2025. https://mindsquare.de/knowhow/kuenstliche-intelligenz/ (Zugriff: 29.06.2025).

Shah, R.; Irpan, A.; Turner, A. M.; Wang, A.; Conmy, A.; Lindner, D.; Brown Cohen, J.; Ho, L.; Nanda, N.; Popa, R. A.; Jain, R.; Greig, R.; Albanie, S.; Emmons, S.; Farquhar, S.; Krier, S.; Rajamanoharan, S.; Bridgers, S.; Ijitoye, T.; Everitt, T.; Krakovna, V.; Varma, V.; Mikulik, V.; Kenton, Z.; Orr, D.; Legg, S.; Goodman, N.; Dafoe, A.; Flynn, F.; Dragan, A.: „An Approach to Technical AGI Safety and Security“. 2025. https://sebastianfarquhar.com/assets/papers/shahApproach2025.pdf?utm_source=chatgpt.com (Letzter Zugriff: 23.11.2025).

Simion, M.; Kelp, C.: „Trustworthy Artificial Intelligence“. In: *Asian Journal of Philosophy* 2 (8), 2023. https://doi.org/10.1007/s44204-023-00063-5 (Letzter Zugriff: 01.01.2026).

Singh, Schelley: „AI's Conversation Leap: Latency Down, Empathy Up". In: *The Times of India. MakeMyTrip*. 03.08.2025. https://timesofindia.indiatimes.com/technology/times-techies/ais-conversation-leap-latency-down-empathy-up/articleshow/123667248.cms (Letzter Zugriff: 10.08.2025).

Singla, Alex; Sukharevsky, Alexander; Yee, Lareina; Chui, Michael: „The State of AI 2025: Agents, Innovation, and Transformation". Mit Beiträgen von Bryce Hall und Tara Balakrishnan. McKinsey Global Institute / QuantumBlack. McKinsey & Company, November 2025. https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/november%202025/the-state-of-ai-2025-agents-innovation_cmyk-v1.pdf (Letzter Zugriff: 30.01.2026).

Soder, Lisa; Smakman, Julia; Dunlop, Connor; Sussman, Oliver: „A Framework for Grading Autonomous Systems by Autonomy Levels". In: *Interface – Journal for Digital Agency & Ethics*, 2025. https://www.interface-eu.org/publications/ai-agent-classification (Letzter Zugriff: 22.01.2026).

Soder, Matthew; Smakman, Rens; Dunlop, Tom; Sussman, Gerald: „An Autonomy-Based Classification: AI Agents, Liability and Lessons from the Automated Vehicles Act". *Policy Brief / Report* 2025.

Thiemann, Ania: „Artificial Intelligence and Competitive Dynamics in Downstream Markets". In: *OECD Roundtables on Competition Policy Papers*, Nr. 331. Paris: OECD Publishing, 14.11.2025. https://www.oecd.org/en/publications/artificial-intelligence-and-competitive-dynamics-in-downstream-markets_ccf0624a-en.html (Letzter Zugriff: 30.01.2026).

Turing, Alan: „Computing Machinery and Intelligence". In: Copeland, Jack (Hg.): *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life: Plus The Secrets of Enigma*. Oxford 2004, 441–464.

Turkstra, Frieso; Nabhani, Sara; Al-Khatib, Khalid: „ARGSBASE: A Multi-Agent Interface for Structured Human–AI Deliberation". In: *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 3: System Demonstrations. 2026, 563–574. https://aclanthology.org/2026.eacl-demo.39.pdf (Letzter Zugriff: 18.03.2026).

Waelen, R. A.: „Rethinking Automation and the Future of Work with Hannah Arendt". In: *Journal of Business Ethics.* 2025. https://doi.org/10.1007/s10551-025-05991-1.

Washington Post: „Masterful Photo Edits Now Just Take a Few Words. Are We Ready for This?". 01.08.2025. https://www.washingtonpost.com/technology/2025/09/01/gemini-flash-nano-banana-ai-photo-editing/ (Letzter Zugriff: 20.08.2025).

Waytz, A.; Epley, N.; Cacioppo, J. T.: „Social Cognition Unbound: Insights into Anthropomorphism and Dehumanization". In: *Current Directions in Psychological Science* 19 (1), 2010, 58–62.

Wei, Fei; Li, Yaliang; Ding, Bolin: „Towards Anthropomorphic Conversational AI Part I: A Practical Framework". In: Cornell University. 2025. https://arxiv.org/abs/2503.04787 (Letzter Zugriff: 20.12.2025).

Weller, C.: „A Robot Has Just Been Granted Citizenship of Saudi Arabia". In: *World Economic Forum*, 26.10.2017. https://www.weforum.org/stories/2017/10/a-robot-has-just-been-granted-citizenship-of-saudi-arabia/ (Letzter Zugriff: 23.12.2025).

Zou, H. P. et al.: „A Call for Collaborative Intelligence: Why Human-Agent Systems Should Precede AI Autonomy". In: arXiv preprint, arXiv:2506.09420 [cs.AI]. 2025. https://arxiv.org/abs/2506.09420 (Letzter Zugriff: 23.01.2026).