Sicherheit als Verstehen

Verstehen als Vertrauen

Theoretische Herleitung und angewandte Evaluation von Gütekriterien zur Erfüllung rechtlicher und ethischer Anforderungen an sichere und interpretierbare KI-Systeme

Universität Osnabrück
Institut für Kognitionswissenschaft

Jonas Niehus
Matrikelnummer: 980240

Masterarbeit

Betreuung:

Dr. T. Thelen (Erstbetreuer)

Dr. U. Meyer (Zweitbetreuer)

Osnabrück, Oktober 2025

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Erkenntnisinteresse

Das Zeitalter der Polykrise Zum Auftakt möchte ich eine Standortbestimmung versuchen. Diese Zeit, die 20er Jahre des 21. Jahrhunderts, unsere Gegenwart, lässt sich analytisch treffend mit dem Ausdruck der *Poly-Krise* charakterisieren. Zählen wir nur einmal für ein Land wie Deutschland, die für große Teile der Bevölkerung spürbar gewordenen Krisen der vergangenen Jahre auf: Klimakrise, pandemische Krise, Wirtschaftskrise, Wohnungskrise, Krise der Ungleichheit, Armutskrise, Wertekrise, Glaubenskrise, Krise der Demokratie und Parteien, um nur einige zu nennen. Diese bilden ein Amalgam sich überlappender und interagierender Krisen. Dem mengt sich nun mit den kriegerischen Auseinandersetzungen, sowie der Destabilisierung der transatlantischen Verhältnisse eine wachsende militärische Bedrohungslage und damit eine Krise der Geopolitik hinzu. Auch eine über 70 Jahre gewachsene und seit über 30 Jahren wiedervereinte Gesellschaft, wie die der Bundesrepublik mit ihrer relativ stabilen Demokratie, hohem Wohlstand und ausgebauten Institutionen, wird durch dieses Krisenamalgam destabilisiert. „Wir sind konfrontiert mit einer noch nie dagewesenen Situation. Das ist [...] keine Übertreibung." (Tooze, 2022, Minute 14:00) Multiple Krisenprozesse, deren Bedingungen sich ineinander verschränken und sich reziprok verstärken, beeinflussen unser aller Leben.[1]

Polykrise als Verstehenskrise Wohl verstanden stellt der Begriff der Polykrise kein analytisches Modell vor, um die multiplen Krisen letztlich doch beschreiben und verstehen zu können, sondern charakterisiert vielmehr einen „kognitiven Schock" (Tooze, 2025) (meine Übersetzung, J.N.), in dem wir die multiplen Krisen epistemisch nicht mehr einfangen können. Es beschreibt nicht die Tatsache, dass wir das Geschehen *noch* nicht verstehen, sondern, dass wir *das* schlichtweg nicht verstehen können. In anderen Worten, die Wirklichkeit ist nicht einfach schwer zu verstehen, weil viele Faktoren wirken, sondern sie ist immer noch komplexer als wir es selbst in unseren besten Modellen, Theorien und

[1] Um es mit den Worten der Weltbank zu resümieren: „(W)e are facing a series of overlapping and interconnected crises that are impacting lives and livelihoods almost everywhere. The combined effects of slow economic growth, rising conflict and fragility, persistent inequality, and extreme weather-related events have sent shockwaves across the globe. High-income economies are showing signs of resilience, but the outlook for low-income economies and fragile countries remains deeply troubling." (The World Bank, 2024)

Gedanken zu verstehen bemüht sind. Es handelt sich demnach um eine „Wissenskrise“[2], der Vielzahl an Tönen, die erklingen, liegt am Ende eben doch nicht *eine* Melodie zugrunde. Und eben aufgrund dieser fehlenden Melodie, der Tatsache, dass wir uns auf die Klänge der Wirklichkeit, wie wir sie kennen und kannten nicht (mehr) verlassen können, ist die Polykrise letztlich eine Krise des Vertrauens schwindelerregenden Ausmaßes. Vertrauen in die Gestaltbarkeit des individuellen und kollektiven Schicksals ist, was wir in diesen Zeiten vermissen und was eine der Quellen ist, aus der sich die gegenwärtigen reaktionären, restaurativen und illiberalen Kräfte speisen.

Informationssicherheit im Zeitalter der Polykrise Fast schon schicksalhaft mengt sich diesem fundamentalem Vertrauensverlust nun noch die stetig wachsende Macht datenverarbeitenden Systeme hinzu (Caspar, 2023; Kirchschläger, 2022). Eine Entwicklung, die als zentraler Treiber des multiplen Krisengeschehens zu benennen ist, ist die Krise der Informationssicherheit und der Öffentlichkeit im Zeitalter der digitalisierten Moderne (BSI, 2024c; Caspar, 2023). Im Kontext der Polykrise ist dabei die enorme Dimension der Risiken durch den Diebstahl, aber vor allem auch durch der Missbrauch von Daten, zu nennen (BSI, 2024c). Böswillige Akteure können die „Macht der Daten“ nutzen, um Menschen am Leib und Leben zu schaden, als sie auch an der grundlegenden Ausübung ihrer Rechte, wie der freien Entfaltung der Persönlichkeit, zu hindern (Caspar, 2023; Dachwitz, 2023; Gille, Meineck & Dachwitz, 2023; Mühlhoff, 2023a, 2023b; Zuboff, 2019). Hierzu sind die Manipulation der Gedanken, die Polarisierung der öffentlichen Meinung, der Diebstahl (sensibler) Informationen, als auch der Ausschluss von lebens- wie grundrechtsrelevanten Gütern[3] als besonders drängende Gefahren zu nennen (BSI, 2024b, 2024c).[4]

Informationssicherheit und künstliche Intelligenz In diesem Umfeld kommen nun lernende, (teil-)autonome, datengetriebene Systeme (DS), sogenannte künstliche Intelligenz, als Gefahrenvektor hinzu (BSI, 2024b; Kirchschläger, 2022). Die schiere Dimension der Datenanalysekapazitäten durch künstliche Intelligenz potenziert die Sicherheitsrisiken. Sie können von böswilligen Akteuren genutzt werden, um die Effektivität der genannten

[2] „For me [...] the polycrisis idea has always been first and foremost about mapping a cognitive shock, rather a well-specified model. The phrase captures what Michael Geyer described to me as a *knowledge crisis*.“ ((Tooze, 2025), m.Ü., J.N.)

[3] Als relevante Güter sind hier zu nennen: Soziale Mobilität, freie Entfaltung der Persönlichkeit, Zugang zu Bildung, Recht auf Gleichbehandlung u.v.m.

[4] Für eine detaillierte Auseinandersetzung mit den verschiedenen datenbasierten Angriffsvektoren auf die informationelle Selbstbestimmung siehe (Caspar, 2023).

Angriffsvektoren dramatisch zu steigern.[5] Exemplarisch sei nur die Manipulation Wahlberechtigter im Superwahljahr 2024 mit KI generierten DeepFakes zu nennen oder sogenannte Information Extraction Attacks, bei denen zum Beispiel personenbezogene Daten aus dem Modell extrahiert werden (siehe auch die von mir erstellte knappe Aufstellung von Angriffsvektoren durch KI (9.2)) (Correctiv, 2024; Crawford, 2021; European Digital Media Observatory, 2024; Meaker, 2023; Sherman, 2024).

Informationssicherheit und Interpretierbarkeit Im Jahr 2025 gelten sogenannte tiefe künstliche neuronale Netze (ANNs)[6] als die leistungsfähigsten datenverarbeitenden Systeme, die aktuell zur Anwendung kommen. In Bezug auf diese Systeme sind wir analog zur Polykrise ebenfalls mit einer epistemischen Krise besonderer Art konfrontiert: Häufig, wenn auch nicht ausschließlich, gehen die Leistungssteigerungen dieser Systeme mit einer Skalierung der Parameter, der Datensätze, der Trainingsepochen und Rechenleistung einher (Kaplan et al., 2020). Hinzu kommen komplexere Lernalgorithmen und Trainingsmethoden, wie zum Beispiel das Reinforcement Learning from Human Feedback (RLHF). Mit diesem Prozess evolvieren zwei wichtige Eigenschaften:

- Durch das Anwachsen der Parameterzahl, der Trainingsdaten und der Kontextdaten in unterschiedlichen Umgebungen wächst die Anzahl der potenziell intransparenten bzw. risikobehafteten Interaktionen zwischen Umgebungsdaten und Modell stetig an (BSI, 2022).

- Modelle ab einer bestimmten Komplexität zeigen bestimmte, nicht-vorhersagbare Eigenschaften (sog. Emergenz (6.2),(8.2.1)) wie zum Beispiel das sogenannte In-Context-Learning (ICL) bei generativen KI-Systemen (Du et al., 2025; Olsson et al., 2022).[7]

Wie wir noch sehen werden, sind diese, während Training und Nutzung evolvierenden

[5] „Es gilt also die Eigenschaften des angelernten Modells zu verifizieren, um die Erklärbarkeit, Robustheit und Sicherheit des KI-Systems (und seiner Reaktionen) nachzuweisen. Insbesondere führt das Fehlverhalten sicherheitskritischer Systeme häufig zu Verlust von Menschenleben und finanzieller Ressourcen. Techniken zum Schutz gegen solche Szenarien sind also für diese Systeme unerlässlich.“ (BSI, 2022)

[6] In dieser Arbeit spreche ich von *künstlichen neuronalen Netzen* oder als Kürzel aufgrund der etablierten Konvention von ANNs (artificial neural networks).

[7] Die jüngere Erforschung des als ICL bekannten Phänomens hat zum Beispiel zutage gefördert, dass Large Language Models nicht einfach nur sogenannte Surface Statistics erlernen, sondern tatsächlich induktive *und* funktional differenzierte Aufmerksamkeitsköpfe evolvieren, womit eine einfache Reduzierung ihrer Komplexität als „stochastische Papageien“ in Zweifel gezogen werden kann (siehe auch (8.2.1)) (Bender et al., 2021; Olsson et al., 2022).

Eigenschaften, der Grund, warum tiefe künstliche neuronale Netze ein für sie charakteristisches Sicherheitsrisiko erzeugen, welches in dieser Arbeit als die Schnittmenge aus Informationssicherheit und Interpretierbarkeit vorgestellt wird (6). Die Gefahr (6.4) besteht darin, dass die genannten Eigenschaften

- zu einem *epistemischen Sicherheitsrisiko* führen, welches daran liegt, dass Menschen Schaden aufgrund der Intransparenz der Modelle erleiden, *ohne* dass es notwendigerweise eine böse Absicht gab (6.4.1),
- von böswilligen Akteuren als Angriffsvektor ausgebeutet werden können (6.4.2) und
- aufgrund der Modellkomplexität und dem damit einhergehenden Verstehensdefizit Sicherheitsrisiken von Akteuren der Informationssicherheit nur schwerlich antizipiert und identifiziert und noch schwerer minimiert werden können (Berghoff, Neu & von Twickel, 2020; BSI, 2021, 2024b, 2025).

Dies ist der Ausgangspunkt für die geläufige Auffassung, dass KI-Systeme basierend auf künstlichen neuronalen Netzen als *Blackboxen* zu charakterisieren sind, das heißt, dass die genaue Beziehung zwischen Input X und Output Y opak ist (Z. C. Lipton, 2016; O. Müller & Lazar, 2024).

Anschaulich können wir Szenarien nehmen, in denen Menschen durch biometrische Identifikationssysteme, Kreditrisikoanalysen oder automatisierte Bewerbungsverfahren der Zugang zu Infrastruktur, Bildung oder beruflichen Chancen verwehrt wird, ohne dass die Gründe für die Ablehnung nachvollziehbar dargelegt werden (können). Alles in allem bedeutet dies, dass unsere Daten und die Daten Dritter der Rohstoff für Prädikationen, Klassifikationen und generierte Medien (Text, Audio usw.) sind, die wir nicht vollständig verstehen können und genau diese mangelnde Erklärbarkeit ist gleichzeitig auch noch ein Sicherheitsrisiko. Im Zusammenhang mit der Informationssicherheit bestimmter KI-Systeme wird der kognitive Schock, die Vertrauenskrise, die die Polykrise charakterisiert, durch eine epistemische Vertrauenskrise des Nichts-Verstehens, des Nichterklärens und Nichtinterpretierens gespiegelt.

1.2 Problemstellung

Da (generative) KI-Systeme zunehmend unser Dasein durchdringen, wird es aus rechtlichen, ethischen und anderweitigen praktischen Gründen immer relevanter, in fraglichen Fällen eine gute Erklärung anbieten und wiederum begründen zu können, was genau unter *guter Erklärung* zu verstehen ist. Denn „wann immer ein KI-System das Leben von Menschen entscheidend beeinflusst, muss es möglich sein, eine geeignete Erklärung für den Entscheidungsprozess des KI-Systems zu erhalten [...].“[8] Hier stellt sich ein zunächst technikinhärentes Dilemma vor. Sobald sich Gesellschaften darauf verständigen Technologien einer bestimmten Komplexität zu nutzen, resultieren hieraus bestimmte Freiheitsgrade, das heißt praktisch nicht-deterministisch vorhersehbare Input-Output-Relation. Folglich ist es kein praktikables Ziel, das Risiko auf Null zu reduzieren. Stattdessen benötigen wir Gütekriterien (7), die hinreichend allgemeine und konkrete Verfahren zur Herstellung von sicheren und vertrauenswürdigen IT-Systemen vorstellen. Sie bieten keine *one-fits-all* Lösung, sondern differenzierte Kriterien, um durch ein undurchsichtiges und gefährliches Terrain zu navigieren. Wie wir sehen werden, gibt es keinen mechanistischen Algorithmus für Interpretierbarkeit und Sicherheit (5.6), sondern im besten Falle pragmatische Heuristiken für die Anwendung von geeigneten Gütekriterien (7.4). Wesentlich wird sich auch die Betrachtung des Einzelfalls als Einzelfall erweisen, das heißt ein jeder fraglicher Fall benötigt eine eigene Methodenkonstellation zur Herstellung eines adäquaten Interpretierbarkeits- und Sicherheitsraums (8.2.2).

Von besonderer Relevanz für diese Arbeit ist die am 01. August 2024 in Kraft getretene KI-Verordnung, die im Zeitverlauf des Abfassens dieser Arbeit sukzessiv zur Anwendung kommt (European Union (EU), 2024).[9] Sie gibt einen Orientierungsrahmen für die in dieser Arbeit diskutierten Gütekriterien. Eine große Herausforderung ist, dass die KI-Verordnung keine spezifischen Benchmarks, Gütekriterien und Praxisleitfäden vorschreibt, was oftmals auch als Lücke in der Anwendung der Verordnung diskutiert wird (Herd et al., 2024).

Damit ist das *Problem und Ziel dieser Arbeit* bestimmt: Die Problemdimension theoretisch,

[8] Und weiter heißt es: „Eine solche Erklärung sollte rechtzeitig erfolgen und auf die jeweilige Sachkenntnis des betroffenen Interessenträgers (z. B. Laie, Regulierungsbehörde oder Forscher) zugeschnitten sein. Darüber hinaus sollten Erläuterungen darüber vorliegen, inwieweit ein KI-System die Entscheidungsprozesse einer Organisation beeinflusst und gestaltet, aber auch über die Entwurfsentscheidungen und die Gründe für die Einführung des Systems (zur Gewährleistung der Transparenz des Geschäftsmodells).“ (Hochrangige Expertengruppe für Künstliche Intelligenz, 2019, S. 22)

[9] Für Details vgl. KI-Verordnung Art. 113 Inkrafttreten und Geltungsbeginn.

rechtlich und ethisch zu erarbeiten und einige Bedingungen für einen Lösungsansatz in Gestalt von Gütekriterien (7) zu erarbeiten, zu evaluieren und zu diskutieren. Die gesamte Arbeit kann als ein Theorie- und Praxisleitfaden für Verantwortliche und Interessierte gelesen werden, welcher weitere Ressourcen an die Hand geben soll, die Blackboxproblematik, wenn auch nicht zu lösen, aber doch zu verstehen und zu moderieren (7.2), (8.1.3), (7.4.2).

Wir können uns eine idealisierte Verantwortliche (2.3) eines Hochrisiko-KI-Systems vorstellen, die XAI-Beauftragte, die die hier entwickelte Argumentation und die daraus resultierenden Gütekriterien in das Risikomanagement (7.2.1) eines Systems einbaut, mit dem *Ziel höchster rechtsethischer Standards*.[10] Ich spreche hier von *rechtsethischen* Standards, um die Doppelausrichtung dieser Arbeit zu markieren. Es geht um Gütekriterien, die sowohl die rechtliche Frage der Compliance behandeln, das heißt dasjenige, wozu wir gesetzlich verpflichtet sind (Legalität), als auch die Frage behandeln, was wir tun sollten, das heißt dasjenige, was wir in komplexen Mensch-Technik-Interaktionsnetzwerken moralisch einander schulden (Legitimität).

1.3 Fragestellung

Möchte man die vorgestellte Problemstellung äußerst schematisch auf drei knappe Fragen reduzieren, dann können die folgenden Hauptfragen, als leitgebend für das Projekt gelten:

1. Welche natürlichsprachigen rechtlichen und ethisch-normativen ***Bedingungen*** und ***Kriterien*** müssen erfüllt sein, um von interpretierbarer und sicherer KI zu sprechen?

2. Unter welchen analytischen ***Bedingungen*** und ***Kriterien*** können wir von (nicht-)interpretierbarer KI sprechen?

3. Welche institutionellen, sowie technisch-formalen ***Bedingungen*** und ***Kriterien*** (sogenannte *Gütekriterien*) sind geeignet um KI Modelle auf ihre Interpretierbarkeit und Sicherheit hin zu prüfen?

Mit der Bearbeitung dieser Fragen bietet die Arbeit gewissermaßen drei Begriffe an: Einen *normativen* Begriff, einen *analytischen* und einen *institutionell-technischen*. Der normative schafft den rechtsethischen Ausgangspunkt, der analytische präzisiert die Problematik

[10] Ich nutze in der Regel geschlechtsneutrale oder weibliche Artikel. Es sollen grundsätzlich immer alle Geschlechter angesprochen sein, obgleich die Abstraktion in dieser Arbeit nicht ohne Risiko ist (8.2.3).

theoretisch und der technische bietet einen eingeschränkten Lösungsweg an. Indem diese drei Schritte in einem einheitlichen sprachlich-formalen Rahmen angeboten werden, sollen sie zusammen ein kohärentes Ganzes ergeben.

2 Methodik

2.1 Interdisziplinärer Ansatz

Die Arbeit umfasst methodisch eine theoretische und eine angewandte Komponente. Dabei sieht sich das Projekt dem interdisziplinärem Geiste der *Cognitive Science* verpflichtet (Stephan & Walter, 2013, S. 23ff.). Die theoretische Komponente bemüht sich mittels verschiedener Disziplinen, vor allem Komplexitätsforschung, Philosophie und Informatik, als auch Jurisprudenz und Soziologie einen Vorschlag zu erarbeiten, wie wir Verstehen und Vertrauen in (hoch-)komplexen Umgebungen und in Interaktion mit Technologien verstehen können. Die Arbeit besitzt dementsprechend einen ontologischen und erkenntnistheoretischen Unterbau. Das heißt, sie skizziert Ansätze einer Theorie, die begreifbar machen soll, wie die Wirklichkeit, einschließlich unserer eigenen Existenz, strukturiert sein muss, damit daraus die beschriebenen Bedingungen der Möglichkeit von (Nicht-)Interpretierbarkeit resultieren. Die angewandte Komponente erarbeitet mittels der Methoden der Informationssicherheit sowie der *Explainable AI*-Forschung und den Bestimmungen des Gesetzgebers einen fragmentarischen Entwurf, wie wir diese theoretische Grundierung als Praxisleitfäden fruchtbar machen können.

Im Bereich Informationssicherheit und Interpretierbarkeit sieht das BSI dringend Handlungsbedarf auf folgenden Ebenen:[11]

1. Entwicklung von Standards, technischen Richtlinien, Prüfkriterien und Prüfmethoden: Derzeit existieren keine hinreichend geeigneten Standards, um die Sicherheit von KI-Systemen für kritische Anwendungskontexte (wie sie z. B. in der Automobil- und Rüstungsindustrie, in der Biometrie, im Gesundheitswesen sowie im Finanz-, IT- und Telekommunikationsbereich vorliegen können) verlässlich zu bewerten und technisch zu prüfen. Auch für weniger kritische Anwendungen fehlen (mit wenigen Ausnahmen) Maßstäbe für die Sicherheit.

[11] Wörtlich zitiert aus (BSI, 2021).

2. Erforschung von wirksamen Gegenmaßnahmen gegen KI-spezifische Angriffe: Die existierenden Maßnahmen für die oben genannten Angriffe sind oft nicht ausreichend. Um einen sicheren und robusten Betrieb von KI-Systemen zu ermöglichen, müssen weitere Gegenmaßnahmen möglichst praxisnah erforscht werden.

3. Erforschung von Methoden der Transparenz und Erklärbarkeit: Die oft mangelhafte Erklärbarkeit von KI-Systemen beeinflusst deren IT-Sicherheit maßgeblich und sorgt für fehlende Akzeptanz der Systeme seitens der Anwender. Es ist daher wichtig, auch die Methoden zur Erklärbarkeit praxisnah weiter zu erforschen (BSI, 2021).

Des Weiteren sieht auch die EU KI-Verordnung enorme Anstrengungen zur Entwicklung von Benchmarks und Gütekriterien zum Zwecke der Minimierung entsprechender Risiken vor (European Union (EU), 2024). Dieser Aufgabenbereich stellt den methodischen Flaschenhals dieser Arbeit vor. Teile der theoretischen und insbesondere die angewandte Komponente dieser Arbeit sind vor allem an der Schnittmenge aus den Ebenen 1 und 3 angesiedelt. Sie analysiert und evaluiert Methoden der Transparenz und Erklärbarkeit (Handlungsfeld 3), zum Zwecke der Diskussion und Evaluation von technischen Prüfkriterien und Prüfmethoden (hier sogenannte Gütekriterien, Handlungsfeld 1).

2.2 Evaluation

Unter *angewandte Evaluation* ist hier das Folgende zu verstehen: Die *formale*, *begriffliche* und *programmbasierte* Analyse der entwickelten Gedankengänge anhand eines relativ kleinen künstlichen neuronalen Netzes, Datensatzes und Anwendungsbeispiels. Diese drei Ebenen werden immer anhand der gleichen Architektur (die hier sogenannte *Standardarchitektur*), des gleichen Datensatzes (den German Credit Data) und des gleichen Fallbeispiels (der Kreditrisikoanalyse) durchgeführt (weitere Details unter (7.5) und (Niehus, 2025)). Unter (7.5.6) können Sie sich schon mit dem Datensatz vertraut machen. Dadurch kann die formale und begriffliche Evaluation auf die programmbasierte zurückbezogen werden und *vice versa*. Die programmbasierte Evaluation dient primär dem Zweck, die hier diskutierten Gütekriterien und den gesamten theoretischen Rahmen anhand eines relativ einfachen Anwendungsbeispiels zu simulieren, zu veranschaulichen, um darauf aufbauend Herausforderungen, Grenzen und Modifikationen zu diskutieren. Dabei wird immer wieder auf den Quellcode referiert. Ich habe den Quellcode, die Klassen und Funktionen für

die Evaluation und Aufbereitung des Datensatzes innerhalb der Notebooks möglichst genau beschrieben, um ein gutes Verständnis zu ermöglichen, auch für Lesende, die nicht versiert sind im Lesen von Quellcode. Hierzu wurde der Quellcode für die Evaluation in 10 Schritten aufgetrennt. Der Quellcode für das künstliche neuronale Netz und das Auswahlprogramm Selection habe ich konventionell durchkommentiert. Zur Reproduktion und Darstellung weiterer technischer Details wurde folgendes Github Repositorium eingerichtet: *https://github.com/PyJonny22/Masterarbeit_Guetekriterien-sichere-und-interpetierbare-Hochrisiko-KI-Systeme.*

2.3 Zielgruppe

Die EU KI-Verordnung adressiert primär Anbietende, Betreibende, Bevollmächtigte, Produktherstellende und Händler:innen von Hochrisiko-KI-Systemen (European Union (EU), 2024, Art. 3). Diese Gruppe können wir für unsere Zwecke zu der Gruppe der Verantwortlichen zusammenfassen, insofern sie ein Hochrisiko System in Verkehr bringen oder in Betrieb nehmen, ein bestehendes im Verkehr befindliches System verändern oder wenn sie wesentliche Änderungen an dem System vornehmen (Herd et al., 2024). Diese Einschränkung ist entscheidend, so sind nicht die Entwickelnden in der Hauptverantwortung, sondern immer die natürliche oder juristische Person, welche das System in Umlauf bringt (Herd et al., 2024). Diese Arbeit richtet sich an eine akademisch interessierte Leserschaft aus dieser Gruppe. Außerdem adressiert sie Verantwortliche in der Wissenschaft, in Forschung und Lehre. Dem hinzuzufügen sind noch die Verantwortlichen in Behörden, wie zum Beispiel dem Bundesamt für Sicherheit in der Informationstechnik (BSI) oder der Bundesnetzagentur (BNetzA). Letzteres soll laut jüngsten Verlautbarungen die zentrale Aufsichtsbehörde für die EU KI-Verordnung in Deutschland werden (Bundesnetzagentur, 2024a, 2024b).

2.4 Vorgehen

Im Folgenden wird einmal kapitelweise das Vorgehen beschrieben, um den Leser:innen eine gute Orientierung zu ermöglichen.

a) Das dritte Kapitel *Taxonomie der KI* (3) beginnt mit einer knappen Definition künstlicher Intelligenz (3.1). Diese Taxonomie wird dann auf die Bestimmungen

zu KI im EU KI Gesetz hin spezifiziert (3.2). Der besondere Ansatz der EU KI-Verordnung in Bezug auf diese Thematik wird herausgearbeitet (3.3).

b) Das vierte Kapitel *Sichere und vertrauenswürdige KI* (4) arbeitet aus der geltenden Gesetzgebung, den Vorschlägen der EU Ethik-Kommission und den Arbeiten einiger renommierter Expert:innen aus Philosophie und Recht ein normatives Verständnis von Sicherheit und Interpretierbarkeit heraus. Aus der Analyse dieser Texte werden sich bereits erste Anforderungen an die Gütekriterien ergeben (4.7).

c) Das fünfte Kapitel *Interpretierbarkeit als epistemische Sicherheit* (5) stellt den Versuch dar, eine interdisziplinäre begriffliche und formale Analyse von Interpretierbarkeit anzubieten und einen einheitlichen analytischen Rahmen für diese Arbeit zu entwickeln. Es geht darum, zu präzisieren, was verstehen in einer komplexen Welt und in Interaktion mit komplexen Maschinen eigentlich bedeutet (5.9).

d) Das sechste Kapitel *Nicht-Interpretierbarkeit als epistemisches Risiko* (6) entwickelt diesen Ansatz zunächst weiter zu einer Ontologie der Komplexität (6.1), um zu präzisieren, wie das Verhältnis zwischen den menschlichen epistemischen Bemühungen und der Komplexität der Wirklichkeit modelliert werden kann. Mit Hilfe dieser Ontologie werden künstlichen neuronalen Netze als komplexe Systeme beschrieben (6.2). Es wird gezeigt, wie aus den (formalen) Eigenschaften dieser Systeme als komplexe Systeme die mangelnde Interpretierbarkeit resultiert (6.3) und damit ihr genuines Sicherheitsrisiko (6.4).

e) Das siebte Kapitel *Gütekriterien* (7) nutzt die Erkenntnisse aus den Kapiteln 2, 3 und 4 und reichert diese um epistemologische und ethische Parameter, als auch Aspekte der Informationssicherheit an, um die Anforderungen und nötigen Eigenschaften von Gütekriterien abzuleiten (7.2). Der analytische Rahmen aus den vorigen Kapiteln wird dann in Form technischer Gütekriterien (7.3) auf künstliche neuronale Netze hin in Form einer idealen Erklärung präzisiert (7.3.1). Die Grenzen der idealen Erklärung werden anhand der Komplexität der Modelle und der Datensätze aufgezeigt und eine eingeschränkte Lösung vorgeschlagen (7.3.4). Daraus werden dann Desiderata für die institutionelle und technische Herstellung von Sicherheit als Verstehen gewonnen (7.4.1) und ein schematischer Ablaufplan der Anwendung dieser Gütekriterien

entwickelt (7.4.2). Dieser wird dann sogleich anhand eines einfachen Modells zu Anschauungszwecken simuliert werden (7.5).

f) Das achte und letzte Kapitel *Sicherheit als epistemisches Vertrauen* (8) wird genutzt, um zunächst einige relevante *Implikationen* dieser Evaluation für das Verfahren der Erklärbarkeit, der Informationssicherheit sowie für Recht und Behörden herauszuarbeiten (8.1). Dann geht es darum, Möglichkeiten und Grenzen in der Entwicklung, der Anwendung und Nutzung im Hinblick auf eine sichere KI und digitale Gesellschaft zu besprechen und schließlich wird der hier entwickelte interdisziplinäre holistische Ansatz und seine formale und technische Implementierung auf seine formalen, methodischen und philosophischen Bedingungen und Grenzen hin befragt (8.2).

2.5 Ziele

Basierend auf dieser Problemskizze und dem beschriebenen Vorgehen können die folgenden Ziele und Beiträge für die 5 Hauptkapitel *4 bis 8* dieser Arbeit formuliert werden:

1. *Kapitel 4:* Ein normatives Konzept sicherer und interpretierbarer KI zu entwickeln und damit den Zusammenhang von Sicherheit und Interpretierbarkeit zu präzisieren.

2. *Kapitel 5:* Begriffliche und formale Bedingungen und Kriterien identifizieren, die einen (technischen) Prozess als interpretierbar charakterisieren.

3. *Kapitel 6:* Bedingungen und Kriterien identifizieren, die ein KI-System als nicht interpretierbar *und* nicht-sicher charakterisieren.

4. *Kapitel 7:* Gütekriterien für sichere *und* interpretierbare KI-Modelle ermitteln, evaluieren und Implikationen diskutieren.

5. *Kapitel 8:* Analysieren und Aufzeigen der Möglichkeiten und Grenzen des institutionell und formal-technischen Ansatzes Sicherheit als Verstehen.

Alles in allem besteht der Beitrag der Arbeit darin, sukzessiv einen ontologisch und rechtsethischen kohärenten, normativen Begriff von Informationssicherheit und Interpretierbarkeit für komplexe KI-Systeme zu gewinnen, diesen analytisch zu präzisieren und in der Form von Gütekriterien formal zu übersetzen, um damit partiell eine technische Überprüfung

auf diesen Maßstab hin zu ermöglichen und dessen Limitationen zu diskutieren. Arbeiten in dem Bereich sind oftmals nicht hinreichend auf die geltende Gesetzgebung abgestimmt und/oder präzisieren den Gegenstand und das Ziel, sowie das zugrundeliegende Verständnis von Interpretierbarkeit nicht ausreichend, was wiederum erhebliche Beschränkungen in der Anwendung bedeutet. Begriffe wie *formale Verifizierung* oder Zielvorstellungen wie *vertrauenswürdige KI* werden oftmals nicht ausreichend mit einem analytisch präzisen *und* rechtsethisch kohärenten Interpretierbarkeitsbegriff untermauert. Es klafft gewissermaßen eine Lücke zwischen den Methoden und der Forschung im Bereich XAI einerseits und deren Spezifikation im Hinblick auf deren Anwendung unter bestimmten rechtlichen, ethischen und epistemischen Desiderata andererseits. Diese Lücke soll hier ein Stück weit geschlossen werden, indem ein einheitlicher begrifflich-formaler Rahmen für alle Schritte angeboten wird und Konstellationen von XAI-Methoden angeboten werden, die bemüht sind der geltenden Gesetzgebung gerecht zu werden. Im Ergebnis haben wir Elemente der KI-Verordnung, wie der Ablaufplan (7.4.2) zeigt, präzisiert. Zu betonen ist noch, dass der normative und analytische Begriff von (nicht-)interpretierbarer KI bereits als Bestandteil der Gütekriterien gelesen werden sollte. Dies ist darin begründet, da der Theorieteil dieser Arbeit der Zielgruppe argumentative Ressourcen geben kann, für bzw. gegen eine bestimmte Methodenkonstellation zu argumentieren. Mit dem theoretischen Fernziel, Sicherheit als Verstehen und Verstehen als Vertrauen auszuzeichnen, weist die Arbeit mit ihren eigenen Methoden über sich hinaus, indem die Bedingungen der Möglichkeiten und damit auch die Grenzen des Ansatzes Sicherheit als Verstehen zu realisieren, thematisiert werden. Wir werden sehen, dass einiges für einen unhintergehbaren menschlichen Beitrag spricht. Technisch-institutionelle Gütekriterien und die menschliche Urteilskraft sind in diesem Rahmen komplementär zu denken. Dann, und nur dann, kann es zu dem seltenen, aber wertvollen Moment kommen, den wir Vertrauen nennen (5.9).

3 Taxonomie künstliche Intelligenz

3.1 Definition

Aufgrund seiner Aktualität und Bedeutung für diese Arbeit kann als Folie für dieses Kapitel die Definition der KI-Verordnung der Europäischen Kommission gelten, welche auch vom Bundesamt für Informationssicherheit (BSI) übernommen wurde (O. Müller &

Lazar, 2024). Dort wird KI umfassend und präzise definiert als

> „ein maschinengestütztes System, das für einen in ***unterschiedlichem Grade autonomen*** Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme ***anpassungsfähig sein kann*** und das ***aus den erhaltenen Eingaben*** für explizite oder implizite Ziele ***ableitet, wie Ausgaben*** wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen ***erstellt werden***, die **physische oder virtuelle Umgebungen beeinflussen** können (Hervorhebungen J.N.).“ (O. Müller & Lazar, 2024), (European Union (EU), 2024, Art. 3)

Diese Definition hebt wesentliche Aspekte hervor und hat den Vorzug gegenüber anderen Definitionen deutlich spezifischer und zugleich hinreichend flexibel zu sein. Die besonderen Merkmale von maschinengestützten Softwaresystemen, die als KI zu charakterisieren sind, ist folglich, dass ein *Input* in einen *Output* transformiert und diese Transformation *teilweise autonom* stattfindet, wobei eine *Anpassungsfähigkeit* an den Input besteht. Dabei sind sowohl Autonomie, als auch Anpassungsfähigkeit keine absoluten und diskreten Eigenschaften, sondern Eigenschaften, die sich in Graden realisieren.[12] Diese Einschränkung ist entscheidend, da durch dieses Kriterium sowohl ein großes Sprachmodell (auch LLM) wie Gemini, als auch ein Expertensystem wie IBMs Watson und auch Web-Technologien wie ein Vorschlag-Algorithmus bei einem Streaming-Dienst erfasst werden. Wir sehen beispielsweise, dass der Grad an Anpassungsfähigkeit und Autonomie im Vergleich zwischen LLMs und Empfehlungssystemen unterschiedlich hoch ist.[13]

3.2 KI in EU-KI Verordnung

Die EU KI-Verordnung orientiert sich auch an der obigen Definition. Um Klarheit zu schaffen, legt der Text Wert darauf, die Eigenschaften und Fähigkeiten von KI-Systemen von „einfacheren herkömmlicheren Softwaresystemen und Programmierungsansätzen ab(zu)grenzen“ (European Union (EU), 2024, (13)). Das wesentliche Merkmal ist, dass KI-Systeme partiell autonom Ableitungen aus gegebenem Input generieren können, wie zum Beispiel Vorhersa-

[12] Der KI-bezogene Autonomiebegriff ist streng von dem Menschenbezogenen Autonomiebegriff zu trennen (siehe auch (4.2), (8.2.1)).

[13] Diese Definition ist für unsere Zwecke ausreichend, die EU hat mittlerweile eine Spezifikation für alle Schlüsselelemente dieser Definition d.i. maschinenbasiert, Autonomie, Anpassungsfähigkeit, Ziele, erstellen oder schließen (im Englischen Gesetzestext Inferencing) von Ausgaben aus den Eingaben vorgelegt (European Commission, 2025).

gen, Inhalte, Empfehlungen oder Entscheidungen, die wiederum ihre physische und digitale Umgebungen beeinflussen können. Diese Fähigkeit prägt sich in Graden der Autonomie aus, die nach der Inbetriebnahme partiell anpassungsfähig sein können, indem sie auf Eingaben aus der "physische(n) oder virtuellen Umgebung"(European Union (EU), 2024, Art. 3) reagieren.

3.3 Risikoklassen

Die EU KI-Verordnung macht von einem risikobasierten Ansatz Gebrauch, das heißt über die obige Definition hinaus werden die verschiedenen Systeme ihren Ableitungsfunktionen und Einsatzfeldern nach in eine aufbauende Hierarchie von Risikoklassen einsortiert. Als grundlegendes Merkmal gilt, dass mit den Risikostufen auch die Anforderungen an Entwickelnde und Betreibende steigt, etwa im Hinblick auf Transparenz und Sicherheit (Herd et al., 2024). Im Folgenden werden die einzelnen Stufen nur knapp vorgestellt und zu jeder Stufe Beispiele gegeben. Da die Hochrisikosysteme hier besonders relevant sind, ist eine ausführliche Darstellung im Appendix ergänzt (9.1).

1. Minimales Risiko: Wenn KI-Systeme unter Einhaltung des allgemein geltenden Rechts, wie zum Beispiel der Datenschutz-Grundverordnung (DSGVO) eingesetzt werden können, ohne dass sie weitere Risiken für Grundrechte darstellen, handelt es sich um ein minimales Risiko. Auch diese Anwendungen bringen Risiken mit sich, doch sind sie theoretisch durch das geltende Recht unter Ausschluss der EU KI-Verordnung bereits ausreichend reguliert, es bedarf also keiner zusätzlichen gesetzlichen Verpflichtungen. Beispiele sind Übersetzungssysteme, Spam-Filter oder Empfehlungssysteme für Medieninhalte (European Union (EU), 2024; Herd et al., 2024).

2. Limitiertes Risiko: Wenn die potenziellen Risiken für die Grundrechte von Menschen als moderat einzustufen sind, dann birgt die Anwendung laut der Verordnung ein limitiertes Risiko. Beispiele sind etwa adaptive Lernsysteme zum Lernen einer Fremdsprache *ohne* Bewertungssysteme für die Schüler, oder Bildgeneratoren, die transparent darstellen, dass es sich um maschinengenerierten Inhalt handelt. Diese Systeme sind bei weitem nicht ohne Risiko, doch betreffen sie nicht direkt die Zuweisung von grundrechtsrelevanten Gütern und Chancen (European Union (EU),

2024; Herd et al., 2024).

3. Hohes Risiko: Wenn die Anwendung grundrechtsrelevante Güter betrifft und potenziell eine Verletzung der Grundrechte darstellt, etwa den Zugang zu medizinischen oder finanziellen Ressourcen regelt, dann gilt sie als Hochrisikosystem. Hierunter fallen zum Beispiel KI gestützte Bewerbungsverfahren oder Prüfsysteme bei der Kreditvergabe (sogenanntes Credit Scoring), als auch Systeme zur Bewertung von Lernergebnissen (Siehe Appendix (9.1)) (European Union (EU), 2024; Herd et al., 2024).

4. Unakzeptables Risiko: Wenn die Anwendung eines KI-Systems ein eindeutiges Risiko für Grundrechte von Menschen darstellt, dann gilt das Risiko als unakzeptabel und diese Anwendungen sind entsprechend verboten. Beispiele sind die automatisierte Bewertung des Verhaltens (sogenanntes Social Scoring) oder die automatisierte Manipulation des Verhaltens, als auch die biometrische Identifikation von Emotionen (sog. Sentiment Analysis) oder die vorausschauende polizeiliche Überwachung (sog. Predictive Policing) (Brennan & Dieterich, 2017; Bröckling, 2019; European Union (EU), 2024; Herd et al., 2024).[14]

5. Systemische Risiken: Mit dem Ausdruck systemische Risiken schafft die EU eine weitere Risikokategorie. Diese wird relevant durch die Verbreitung sogenannter Foundation Models, also Grundlagenmodelle. Ihr besonders Charakteristikum ist, dass sie gerade nicht für eine spezifizierbare Aufgabe entwickelt wurden (sogenannte narrow AI), wie Übersetzung oder Code Assistenz, sondern für eine Vielzahl unterschiedlichster Anwendungen fein abgestimmt werden können und dabei sehr leistungsfähig sind (sog. General Purpose AI). Generative KI-Systeme wie Transformer Sprachmodelle sind als ein Beispiel zu nennen. Folglich können Modelle dieser Art auch nicht problemlos einer Risikokategorie zugeordnet werden, da sie je nach Feinabstimmung

[14]Es ist wichtig darauf hinzuweisen, dass diese Bestimmungen zwar formal korrekt sind, doch die empirische Realität noch Zweifel aufkommen lässt. Dies hat auch damit zu tun, dass es nicht restlos geklärt ist, welche Anwendungen konkret unter den Begriff Social Scoring oder Predictive Policing fallen. De facto werden Anwendungen, die als solche zu klassifizieren sind, sowohl von privaten, als auch staatlichen Stellen eingesetzt. Beispielsweise sei hier nur an den aktuellen Referentenentwurf des Bundesinnenministeriums erinnert, welcher vorsieht, bundesweit den Einsatz von Spyware der umstrittenen Softwarefirma Palantir auszudehnen (Deutschlandfunk, 2025; L. Kaiser, 2017; Rudl, 2018). Darüber hinaus ist die Überwachungssoftware von Palantir schon seit Langem in unterschiedlichen Bundesländern im Einsatz oder wird von Behörden getestet (Netzpolitik.org, 2024a, 2025a, 2025b).

unterschiedlichen Kategorien zuzuordnen ist. Diesem Charakteristikum soll mit dieser Risikokategorie begegnet werden (European Commission, 2024b; European Union (EU), 2024; Herd et al., 2024; Leisegang, 2024; Meineck, Köver & Leisegang, 2024).

3.4 Eingrenzung

KI-Systeme, welche ein hohes bis unakzeptables Risiko ausprägen können und gleichzeitig kein systemisches Risiko bergen, sind nach obiger Definition auch als enge oder spezifische KI (engl. narrow AI) zu klassifizieren. Die Beispiele, die unter hohes Risiko (3) gegeben wurden, stellen solche Systeme dar. Charakteristisch ist, dass per Definition von diesen Systemen ein hohes oder unakzeptables Risiko in einem Anwendungsbereich, aber ohne menschliche Modifikation des Systems keine Gefahr in einem anderem Bereich zu erwarten ist (Herd et al., 2024). Dies unterscheidet sie von den Systemen mit systemischem Risiko, die das Potential haben, abhängig vom *Fine-Tuning* ein hohes bzw. unakzeptables Risiko in unterschiedlichsten Domänen auszuprägen. Sowohl spezifische KI-Systeme mit hohem Risiko, als auch solche mit systemischen Risiko sind stetig mehr in Verbreitung. Um eine handhabbare Ausgangslage zu haben und den Umfang dieser Arbeit nicht zu sprengen, fokussiere ich mich hier tendenziell auf spezifische Hochrisiko-KI-Systeme. Diese gelten gewissermaßen als Bezugsrahmen für die gesamte Arbeit, ihre Theorie, die genannten Beispiele und die Evaluation. Dennoch wird an der ein oder anderen Stelle und insbesondere auch in der Diskussion erwogen, wie der hier vorgestellte Rahmen auf die Modelle mit systemischem Risiko erweitert werden kann (8.2.1).

4 Sichere und vertrauenswürdige künstliche Intelligenz

In dieser Überlegung sollen aus zentralen Gesetzestexten und (rechts-)ethischen Beiträgen sowie Quellen aus der Informationssicherheit ein normativer Ausgangspunkt gewonnen werden. Es geht um die Frage, welchen Zielzustand bei der Entwicklung, Testung und Anwendung von KI-Systemen wir anstreben *sollten*. Welche Güter sollen von *Güte*kriterien geschützt werden und sind schützenswert?[15] Dabei gibt es zahlreiche, relevante Dimen-

[15]Die folgenden Ausführungen erheben den Anspruch mit den zitierten Texten konsistent zu sein, dennoch ist sich der Autor bewusst, dass eine solche normative Präsentation immer auch eine Interpretation der dargelegten Texte ist. Es geht um eine Interpretation, wie wir diese rechtsethischen Güter verstehen können.

sionen, die im Kontext von KI herausgearbeitet werden könnten. Doch hier liegt der Fokus insbesondere darauf die Schnittstelle zwischen Sicherheit und Interpretierbarkeit zu explizieren. Dabei tritt immer auch ein bestimmtes rechtsethisches Menschenbild zutage, ein Bild davon, wer wir sind und wer wir sein wollen bzw. sein sollen. Die unterschiedlichen Stränge werden zu einem anthropologischen Ausgangspunkt zusammengeführt, welcher durchaus normatives Potenzial hegt. Die zentralen Textgrundlagen für diesen Beitrag lauten in Reihenfolge ihrer Veröffentlichung:[16]

- Grundgesetz für die Bundesrepublik Deutschland (1949)
- Charta der Grundrechte der Europäischen Union (2009)
- Datenschutz-Grundverordnung der Europäischen Union (2018)
- Responsible Artificial Intelligence (2019)
- Ethik-Leitlinien für eine Vertrauenwürdige KI (2019)
- On Artificial Intelligence - A European approach to excellence and trust (2020)
- Ethische KI? Datenbasierte Systeme (DS) mit Ethik (2022)
- Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz (2022)
- Verordnung der EU zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (2024)
- Whitepaper Transparenz KI-Systeme (2024)

Theoretische Schützenhilfe erhalten wir von einigen umfassenderen Werken aus der theoretischen Philosophie, von denen unter Anderem vor allem die folgenden eine gewichtige Rolle spielen:

- Grundlegung zur Metaphysik der Sitten (Kant, 1785)
- Im Reich der Freiheit: Hegels Theorie autonomer Vernunft (Knappig, 2013)
- A Spirit of Trust: A Reading of Hegel's Phenomenology (Brandom, 2019)

[16]Jahr des Erscheinens bzw. Inkrafttretens in Klammern dahinter (XXXX).

4.1 Technik als Instrument

Der angestrebte Zielzustand, welcher sich in diesen Texten ausdrückt, lässt sich mit dem Titel *KI als Werkzeug* versehen. Dies gilt für alle relevanten Dimensionen: Datenbeschaffung, Datenaufbereitung- und -verarbeitung, als auch Beschreibung der Funktionen und Zwecke eines KI-Systems sowie deren ständiger Gebrauch. Werkzeuge werden im engen Sinne als technische Artefakte verstanden, die zweckgebunden in einem überschaubaren Einsatzfeld menschlichen Interessen dienen (Heidegger, 1954). Dieses instrumentelle Verständnis zeigt sich zum Beispiel in der Datenschutz-Grundverordnung in den Passagen, die betonen, dass personenbezogene Daten zweckgebunden und für Betroffene nachvollziehbar verarbeitet werden müssen (European Union (EU), 2016, Art. 5, Art. 12) oder in den Passagen der KI-Verordnung, die die Zweckbindung von Algorithmen und die Robustheit der Systeme betreffen (European Union (EU), 2024, (27)). Robustheit drückt die Erwartung an Technologien aus, nicht unerwartet bzw. aus den falschen Gründen vom vorgesehenen Zweck abzuweichen (European Union (EU), 2024, (27)).[17] Beispielsweise sollten zwei in allen relevanten Werten identische Profile als Eingabedatum für ein Kreditanalysesystem nicht zu unterschiedlichen Risikoprognosen führen. Im Hintergrund spricht hier der Gleichheitsgrundsatz als Rechtsstaatsprinzip, wesentlich Gleiches gleich zu behandeln und wesentlich Ungleiches ungleich (GG, 1949, Art. 3(1)). Dies schließt an die Forderung an Entwickelnde an, tendenziell transparente Systeme zu entwickeln. „Um Bedenken hinsichtlich der Undurchsichtigkeit und Komplexität bestimmter KI-Systeme auszuräumen und die Betreibende bei der Erfüllung ihrer Pflichten gemäß dieser Verordnung zu unterstützen, sollte für Hochrisikosysteme Transparenz vorgeschrieben werden, bevor sie in Verkehr gebracht oder in Betrieb genommen werden.“ (European Union (EU), 2024, (72)) Das Gesetz beschreibt Technologien oft als Instrumente, deren Zwecke begrenzt sind und diese Grenzen sollen von Betreibenden transparent gemacht werden.[18] Insbesondere in der KI-Verordnung rückt Transparenz ins Zentrum. Transparenz umfasst die Nachvollziehbarkeit von Entscheidungen und die Bereitstellung relevanter Informationen über die Fähigkeiten,

[17] „Technische Robustheit und Sicherheit bedeutet, dass KI-Systeme so entwickelt und verwendet werden, dass sie im Fall von Schwierigkeiten robust sind und widerstandsfähig gegen Versuche, die Verwendung oder Leistung des KI-Systems so zu verändern, dass dadurch die unrechtmäßige Verwendung durch Dritte ermöglicht wird, und dass ferner unbeabsichtigte Schäden minimiert werden.“ (European Union (EU), 2024, (27))

[18] „Nach den Leitlinien der hochrangigen Expertengruppe bedeutet *menschliches Handeln und menschliche Aufsicht*, dass ein KI-System entwickelt und als Instrument verwendet wird, das den Menschen dient, die Menschenwürde und die persönliche Autonomie achtet und so funktioniert, dass es von Menschen angemessen kontrolliert und überwacht werden kann.“ (European Union (EU), 2024, (27))

Einschränkungen und Entscheidungslogiken eines Systems. Auch die Rückverfolgbarkeit und Erklärbarkeit von Prozessen und Daten, welche Entscheidungen beeinflussen, sollen zum Beispiel durch Dokumentationspflichten realisiert werden (European Union (EU), 2024, (27)).

4.2 Würde

Im Hintergrund drückt sich dabei ein noch wesentlicheres rechtsethisches Gut aus, nämlich die Unveräußerlichkeit der Menschenwürde.[19] Diese wurde in der BRD historisch in der Debatte um Artikel 1 des Grundgesetzes durch die *Objektformel* präzisiert. In Kürze besagt diese, dass sich eine Behandlung verbietet, die einen Menschen zum bloßen „Objekt staatlichen Handelns“ reduziert (Epping, Lenz & Leydecker, 2024, S. 346). Mit dem hier ausgedrückten Instrumentalisierungsverbot bekommt die Menschenwürde als ethisches Gut einen positiven Gehalt. Treffend wird dies von Immanuel Kant ausformuliert, auf den die begriffliche Präzisierung der Menschenwürde oftmals zurückgeführt wird. Er bietet ein anschauliches Bild, um diesen Gedanken verständlich zu machen (Kant, 2000, S. 68). Kant legt dar, dass es letztlich nur zwei Kategorien von Werten gebe: *Preis* oder *Würde*. Kant präzisiert den Ausdruck Preis dahingehend, dass es im Prinzip ein Äquivalent zu diesem Gegenstand gibt. Nach Kant haben alle Gegenstände, wie Güter des täglichen Bedarfs letztlich einen Preis. Auf die Objektformel gewendet ließe sich sagen, sie sind *bloße* Objekte. Die Würde lässt sich nun in einem ersten Anlauf negativ so bestimmen, dass wenn etwas keinen Preis hat, dann muss es *ex hypothesi* folglich eine Würde haben. Bildlich gesprochen bedeutet dies, dass die Würde etwas ist, das mit nichts anderem aufgewogen werden kann. Wenn etwas eine Würde hat, dann darf es folglich nicht nur als Mittel zum Zweck genutzt werden. Aus dieser Bestimmung leitet sich die auf die Menschenwürde spezifizierte Formulierung seines kategorischen Imperativs ab: „Handle so, dass du die Menschheit sowohl in deiner Person, als in der Person eines jeden anderen jederzeit zugleich als Zweck, niemals bloß als Mittel brauchst.“ (Kant, 2000, S. 61) Nach Kant haben nun Menschen *als* Vernunftwesen eine Würde.[20] Die Menschenwürde wurde durch den Gesetzgeber und die

[19] „Menschenwürde kann [...] als der Eigenwert des Menschen beschrieben werden, der dem Menschen kraft seines Personseins zukommt. Jeder Mensch besitzt – mit den Worten des BVerfG – als Person diese Würde, ohne Rücksicht auf seine Eigenschaften, seinen körperlichen oder geistigen Zustand, seine Leistungen und seinen sozialen Status. Sie kann keinem Menschen genommen werden und geht auch durch *unwürdiges* Verhalten nicht verloren.“ (Epping, Lenz & Leydecker, 2024, S. 346)

[20] Ich klammere hier die zahlreichen Probleme aus, welche mit Kants Würdebegriff einhergehen, insbesondere der Vorwurf, dass aus seinem Prämissenrahmen folgte, dass ausschließlich erwachsene, vernünftige

Rechtsprechung des Bundesverfassungsgericht dahingehend erweitert und präzisiert, als dass es sich bei der Menschenwürde um den unveräußerlichen „Eigenwert des Menschen“ handelt (Epping, Lenz & Leydecker, 2024, S. 346). Das heißt, die Menschenwürde kann, anders als andere Güter wie etwa Abschlüsse, ein Anstellungsverhältnis oder Vermögen *nicht erworben* und *nicht verloren* werden. „Jeder Mensch besitzt– mit den Worten des BVerfG– als Person diese Würde, ohne Rücksicht auf seine Eigenschaften, seinen körperlichen oder geistigen Zustand, seine Leistungen und seinen sozialen Status. Sie kann keinem Menschen genommen werden und geht auch durch *unwürdiges* Verhalten nicht verloren.“ (Epping, Lenz & Leydecker, 2024, S. 346) Auf die Technik angewandt leitet sich aus dem Instrumentalisierungsverbot das oben beschriebene Technikverständnis ab (für Details siehe (Becker, 1996; Dreier, Epping & Lenz, 2024)). Denn wenn technische Artefakte nicht als Instrumente verstanden würden, ließe sich nicht ausschließen, dass Menschen zu Instrumenten dieser Artefakte würden. Damit ist nicht gesagt, dass die real existierenden technischen Artefakte tatsächlich diesem Maßstab gerecht werden (Deutscher Ethikrat, 2023; Mühlhoff, 2023a; Winner, 1980). Zugespitzt formuliert könnte man sagen, dass der liberale Rechtsstaat bis zu einem gewissen Grad in seinem normativen Selbstverständnis und seiner Arbeitsweise auf ein instrumentelles Technikverständnis verpflichtet ist (siehe auch (8.2.7)). Dies ist die normative Grundierung von Beschränkungen, die zum Beispiel automatisierte Entscheidungen mit Personenbezug betreffen.

4.3 Autonomie

Die Menschenwürde beschränkt die Instrumentalisierung zum Zwecke anderer Menschen oder Maschinen. Dies öffnet die Frage, was eine nicht-instrumentelle Zweckgebung ist, nach der sich Menschen richten können. Wenn der Mensch seine Zwecke nicht ausschließlich von außen erhält (Instrumentalisierung), sondern sich diese selber gibt, dann sei dies moralisch gut und verdient den Namen Autonomie. Sie beschreibt die *Selbst-Gesetzgebung* vernünftiger Wesen (Deutscher Ethikrat, 2023; Kant, 2000). Dabei verstehe ich unter Autonomie, dass der betroffene Mensch vernünftiger Weise unter der Bedingung der Kenntnis aller relevanten Informationen, diesem Zweck zustimmen kann. Nun stellt sich hier unmittelbar das Problem ein, dass eine gewisse Instrumentalisierung der Anwendung von Technologien

Menschen eine Würde hätten und welche Implikationen dies für Menschen mit Behinderung und Kinderethik hat (siehe zum Beispiel (Henning, 2016)).

inhärent zu sein scheint. Wenn wir die hier behandelten Hochrisiko-KI-Systeme in den Blick nehmen, wie etwa KI gestützte Kreditrisikobewertung, medizinische Studien oder Bewerbungsverfahren, dann ist es substanziell für diese Anwendungen, dass Menschen, ihre persönlichen Bedürfnisse sowie personenbezogen Informationen und Rechte anteilig Gegenstand einer datengestützten Instrumentalisierung werden. Dieses Dilemma kann man auch als die fundamentale Spannung verstehen, welche sich durch die Digitalgesetzgebung zieht, die immer Ausdruck zweier Perspektiven ist. Einerseits diejenige, die eine instrumentelle Anwendung ermöglichen will, die oft von optimistischen Zukunftsszenarien von Prosperität und Fortschritt begleitet wird und andererseits diejenige Perspektive, die rechtsethische Güter zu schützen beansprucht und damit auch das Instrumentalisierungsverbot respektive das Ziel einer autonomen Lebensführung (beispielsweise (European Union (EU), 2024, (1) und (5))). Der Gesetzgeber, aber auch alle weiteren Stakeholder wie Behörden und Betreibende, haben die anspruchsvolle Aufgabe, diese Perspektiven miteinander zu versöhnen. Dem Gesetzgeber kommt dabei die Aufgabe zu, Verfahren zu etablieren, die bei der Zweckbestimmung durch andere, eine autonome Entscheidung ermöglichen, zum Beispiel durch das Bundesdatenschutzgesetz, das Arbeitsrecht und so weiter (Epping, Lenz & Leydecker, 2024). Ein konkretes Beispiel dafür bietet das wiederum idealisierte Szenario einer pharmazeutischen Studie gemäß Arzneimittelverordnung, Datenschutz-Grundverordnung und GCP-Richtlinien zu einem neuen Diabetes-Medikament. Zunächst liegt hier eindeutig ein Instrumentalisierungsverhältnis vor, da die Teilnehmenden der Studie Mittel zum Zweck der Medikamentenentwicklung sind. Jedoch sieht der Rechtsstaat spezifische Verfahren vor, um Autonomie zu wahren bzw. herzustellen: Teilnehmende werden umfassend über Zielsetzung, mögliche Risiken, Nebenwirkungen und Alternativen der Studie aufgeklärt. Zudem erhalten sie transparente Informationen darüber, wie ihre persönlichen Daten genutzt werden und in welcher Form Ergebnisse veröffentlicht werden sollen. Mit diesen Informationen soll jede teilnehmende Person prüfen können, ob sie die Zwecke und Bedingungen der Studie für angemessen und vertretbar hält (World Medical Association, 2024). Wenn die Proband:innen nach dieser transparenten und informierten Aufklärung freiwillig zustimmen, können wir unter Umständen ein autonomes Selbstverhältnis attestieren.[21] Und in diesem Sinne wäre dann der Schutz der Menschenwürde

[21]Der Autor ist sich bewusst, dass dieses Beispiel ein rechtlich-idealisiertes Szenario ist und die Praxis durchaus von Missbrauch dieser Richtlinien und Manipulation von Probanden geprägt ist. Es geht an dieser Stelle auch nur darum, dass ideelle Verfahren zu veranschaulichen.

gewahrt. Dieses Autonomieverständnis realisiert sich in der Digitalgesetzgebung durch die diversen Bestimmungen zur Dokumentation, Information und Aufsichtspflicht in der KI-Verordnung oder dem Prinzip zur Integrität und Vertraulichkeit personenbezogener Daten in der DSGVO (European Union (EU), 2016, 2024). Insbesondere das Grundrecht auf informationelle Selbstbestimmung, wie es im Grundsatzurteil des Bundesverfassungsgerichts zur Volkszählung aus der Taufe gehoben wurde, lebt von dem zentralen Stellenwert der Autonomie (Epping, Lenz & Leydecker, 2024, S. 362ff.), (*Beschluss des Bundesverfassungsgerichts: "Volkszählungsurteil" (Recht auf informationelle Selbstbestimmung)*, 1983). In diesem Verständnis habe ich als Individuum ein umfassendes Bestimmungsverhältnis über die mich betreffenden Informationen. Ich soll bestimmen können, ob und wie Informationen, welche Eigenschaften über meine Persönlichkeit codieren, von anderen Personen und Systemen genutzt und verarbeitet werden. Die Kategorie der menschlichen Aufsicht, die sich in allen relevanten Texten zum Thema wiederfindet, unterstreicht nochmal die Bedeutung der Autonomie in diesem Zusammenhang. „Hochrisiko-KI-Systeme sollten so gestaltet und entwickelt werden, dass natürliche Personen ihre Funktionsweise überwachen und sicherstellen können, dass sie bestimmungsgemäß verwendet werden und dass ihre Auswirkungen während des Lebenszyklus des Systems berücksichtigt werden." (European Union (EU), 2024, (73)) So heißt es auch in den Leitlinien der Expertengruppe „menschliches Handeln und menschliche Aufsicht (bedeutet), dass ein KI-System entwickelt und *als Instrument* (meine Hervorhebung, J.N.) verwendet wird, das den Menschen dient, die Menschenwürde und die persönliche Autonomie achtet und so funktioniert, dass es von Menschen angemessen kontrolliert und überwacht werden kann." (Deutscher Ethikrat, 2023) Alles in allem können wir aus den rechtlichen und ethischen Schlüsseltexten ein autonomiebasiertes Freiheitsverständnis ableiten, welches weit darüber hinausgeht Freiheit mit der Abwesenheit von Zwang zu identifizieren und diese mit einer eng umrissene Sphäre des Privaten auszustatten.

4.4 Freiheit als autonome Vernunft

Diese rechtsethische Bestimmung des Menschen, dieses Menschenbild, ist normativ, da hier die Autonomie zur Norm erhoben wird. „Menschen sind befähigt zur Handlungsurheberschaft und somit zur Autorschaft ihres Lebens. Sie sind frei und tragen daher

Verantwortung für die Gestaltung ihres Handelns.“ (Deutscher Ethikrat, 2023, S. 146)[22] Und dies geht mit einer elementaren anthropologischen Dimension einher: „*Die Affektion durch Gründe*“ (Deutscher Ethikrat, 2023, S. 146). Es gibt eine (potenziell unendliche) Vielzahl an Faktoren, die Menschen affizieren und menschliche Handlungen bedingen (Triebe, Instinkte, Zwänge, Emotionen, Überzeugungssysteme u.v.m.).[23] Doch das Menschenbild des modernen Rechts- und Verfassungsstaats und damit auch ganz wesentlich das der Republik, ist konstitutiv auf die Prämisse angewiesen, dass menschliche Handlungen mindestens anteilig durch Gründe *bedingt* und *erklärt* werden können (Knappik, 2013, S. 93ff.), (Deutscher Ethikrat, 2023, S. 146f.). Dabei ist eine zentrale Erkenntnis der Philosophie, dass Gründe in begrifflich vermittelten Diskursen normative Entitäten sind (Deutscher Ethikrat, 2023, S.146f.). Eine bestimmte Begriffsverwendung (*das x dort vorne ist rot und nicht grün*) wirkt im *Spiel des Gebens und Verlangens von Gründen* sanktionierend und normierend (Sellars, 1999, S. 66). Gründe etablieren eine Norm der Wahrheit.[24] Gründe sprechen für oder gegen Überzeugungen oder Handlungen.[25] Menschen seien demnach prinzipiell in der Lage durch rationale Reflexion unter Einbeziehung der inferentiellen Beziehungen die Gründe ihres Denkens und Handelns zu erörtern, abzuwägen und zu evaluieren, um sich *praktische Identitäten* anzueignen oder abzulehnen (Knappik, 2013, S. 102ff.). Diese praktischen Identitäten sind in Kürze so zu definieren, dass sie Prädikate darstellen, unter denen ein Selbst sich potenziell wertschätzen kann. Beispiele für solche praktischen Identitäten sind etwa das Dozent:in-Sein oder das Handwerker:in-Sein, ebenso das Muslim:in-Sein oder das Selbstverständnis, eine gute KI-Entwickler:in zu sein, die die hier beschriebenen Gütekriterien anwendet und wertschätzt (Knappik, 2013, S. 11f.). Durch die Wertschätzung dieser Identitäten werden diese zu Quellen von Gründen für unsere Lebensführung (Knappik, 2013, S. 113f.), (Deutscher Ethikrat, 2023). Durch diesen

[22] „Das Grundgesetz sieht die freie menschliche Persönlichkeit und ihre Würde als höchsten Rechtswert an. So hat es folgerecht in Art. 4 Abs. 1 die Freiheit des Gewissens und seiner Entscheidungen, in denen sich die autonome sittliche Persönlichkeit unmittelbar ausspricht, als ünverletzlichänerkannt.“ (Epping, Lenz & Leydecker, 2024, S. 347)

[23] Wobei der Ausdruck Handlung hier noch unterminologisch verstanden werden sollte und nicht wie oftmals als eine Aktion mit starker Urheberschaft von Seiten des Akteurs.

[24] „Wissen ist also nur in einer apologetischen Dimension möglich [...]. Dies bedeutet aber, dass es nur im Kontext sozialer Praktiken angemeldet werden kann, die festlegen, was als Rechtfertigung, was als Grund gelten soll.“ (Gabriel, 2016a, S. 296), (Brandom, 2019)

[25] „Praktische Gründe sprechen für Handlungen, sie sind per se normativ, nicht erst über den Umweg eigener Wünsche. Ein Grund spricht dafür, das zu tun, was diesen Grund erfüllt, wenn nicht andere Gründe dem entgegenstehen. Theoretische Gründe sprechen für Überzeugungen; auch diese sind normativ. In der Regel gibt es Gründe, das eine zu tun und das andere zu lassen, die gegeneinander abgewogen werden müssen. Der Konflikt von Gründen zwingt dann zur Abwägung und zur Systematisierung dieser Abwägung in Gestalt ethischer Theoriebildung.“ (Deutscher Ethikrat, 2023, S. 146)

Prozess kann sich der Mensch mit den Bedingungen des eigenen Denkens und Handelns identifizieren bzw. sich von diesen entfremden (Knappik, 2013, 103ff.). Unsere Handlungen sind demnach nicht einfach unwillkürliche, zwanghafte oder mechanische Körperbewegungen, sondern mindestens anteilig für eine rationale Ausrichtung an Normen offen (Knappik, 2013, S. 111ff.). Dadurch, dass diese Aneignung und Ausrichtung eine gewisse zeitliche Persistenz und logische Kohärenz hat, kann sich der Mensch über verschiedene Kontexte hinweg *re*-identifizieren und gegebenenfalls korrigieren und somit kommt es zur Möglichkeit der Ausprägung eines partiell autonomen Selbst, der Person. „Die rationalen Strukturen, die die Transformation und Aneignung von Willensinhalten ermöglichen, sind konstitutiv für uns selbst– wir selbst sind die Vernunft und müssen uns nicht erst mit ihr identifizieren." (Knappik, 2013, S. 104), (Deutscher Ethikrat, 2023) Im idealisierten Szenario der Ausbildung eines autonomen Selbst als „rationale Persistenz" eignen sich Menschen in einem komplexen Prozess, den wir als *rationale Transformation* fassen können, die Gründe für ihre Lebensweise als die eigenen Gründe an. „Indem wir entscheiden, wie wir handeln, entscheiden wir, in welcher Weise wir kausal wirksam werden, mithin [...] auch, wer wir sind." (Knappik, 2013, S. 111)[26] Diese Bestimmungen können wir interessanterweise in Einklang lesen mit den Grundprinzipien der Datenschutz-Grundverordnung, wie Zweckbindung, Vertraulichkeit und Integrität, als auch insbesondere der Rechtsprechung des Europäischen Gerichtshofs und des Bundesverfassungsgerichts. Diese grenzen nicht einfach eine private Sphäre ein, die es zu schützen gilt, sondern versuchen eine umfassende informationelle und damit personelle Integrität herzustellen. Bringen wir all dies nochmal mit Knappig auf den Punkt, der diese Verhältnisse in seiner großen Hegel Studie systematisch ausgearbeitet hat: „Selbst, Freiheit und Gründe sind Aspekte ein und derselben Sache, die nur in ihrer Wechselbeziehung überhaupt existieren können: Es gibt kein Selbst ohne Freiheit und Gründe, keine Freiheit ohne Gründe und Selbst, und keine Gründe ohne Selbst und Freiheit." (Knappik, 2013, S. 103) Eine wichtige anthropologische Beobachtung an dieser Stelle ist, dass praktische Identitäten im Laufe der Menschwerdung bzw. Persönlichkeitsentwicklung Teil eines partiell impliziten, als auch expliziten Selbstbildes werden. In der jüngeren Philosophie des *Neoexistenzialismus* wurde für diesen Prozess der Name *Geist* rehabilitiert (Gabriel, 2020b). *Geist* beschreibt den Sachverhalt, dass menschliches Denken und Handeln

[26] Im Kontext der Willensfreiheitsdebatte schreibt Meyer „[w]enn Entscheidungen wirklich meine eigenen sein sollen, dann müssen sie ferner ein über die Zeit hinweg verständliches Muster ergeben – sie müssen sozusagen narrativ nachvollziehbar sein." (Meyer, 2024, S. 24)

sich (partiell) an einem individuellen und kollektiven Selbstbild darüber orientiert, wer oder was der Mensch ist (Gabriel, 2020a, S. 229ff.). Und als Teil dieses Selbstbildes werden die praktischen Identitäten wirksam als Gründe für die Lebensführung (Knappik, 2013, S. 112f.), (Korsgaard, 2009, S. 17ff.). Sogenannte objektstufige Anthropologien würden daran scheitern, dass sie den Menschen anhand spezifischer Menschenbilder zu bestimmen versuchen, wie zum Beispiel *homo oeconomicus, homo ludens, homo faber* und so weiter. Dem hält der Neoexistenzialismus entgegen, dass es nicht der Wettbewerb verschiedener Menschenbilder ist, der die richtige Anthropologie ermitteln soll, sondern die formale Fähigkeit, sich ein Menschenbild zu entwerfen und an diesem zu orientieren, sei selbst die akkurate Anthropologie. Der Neoexistenzialismus vertritt die These, dass Geist die formale Invariante beschreibt, die alle Menschen teilen (Gabriel, 2020a, §6, §7). Wenn diese Anthropologie erstens realistisch und zweitens wünschenswert ist, dann folgt aus dieser Besprechung bereits ein entscheidendes Desideratum für das Problem der Interpretierbarkeit. Erklärungen sind Gründe, denen Menschen zustimmen oder die sie ablehnen können. Zum Prozess einer idealen, erfolgreichen Erklärung in einem fraglichen Fall gehört damit, dass die Person diese Gründe verstehen und anerkennen und diese in ihr Selbstbild integrieren kann. Das Ziel einer gelungenen Mensch-Technik-Interaktion müsste *idealiter* sein, dass unsere Autonomie und Autorschaft, wie sie vom deutschen Ethikrat gefasst wird, im Prozess einer solchen Aneignung, dessen Resultat ein mündiges (bzw. mündigeres) Selbst sein sollte, erweitert und nicht beschränkt wird (Deutscher Ethikrat, 2023, S. 178).[27]

4.5 Diskursive und sozialisierte Vernunft

Die *rationale Transformation* zur autonomen Vernunft mag bis hierhin wie epistemologisch privater Prozess anmuten. Das heißt, dass der Mensch oder wie wir im Folgenden sagen werden, das epistemologische Subjekt S, sich praktische Identitäten als Gründe des Denkens und Handelns primär privat aneignet, bevor diese Gegenstand gesellschaftlicher Vermittlung werden, zum Beispiel im Bemühen eines Verantwortlichen die Entscheidungslogik eines vermeintlichen Blackbox Modells zu verstehen. Die Erkenntnis der Welt ist primär logisch privat und die soziale Vermittlung im Spiel des Gebens und Verlangens von Gründen sekundär. Die sozialontolgische Tiefe dieses Problems wird uns in der Diskussion

[27]Wie wir sehen werden (6.1), ist die Annahme, dass Abstraktion in der Form von abstrakten Gründen handlungswirksam wird konsistent und wird ausdrücklich gestützt von der naturwissenschaftlichen Komplexitätsforschung nach Ellis (Ellis, 2012).

noch wieder begegnen (8.2.5). Wichtig ist hier zunächst, diesem Modell die Erkenntnisse einer (im Einzelnen freilich ganz unterschiedlich gelagerten) Strömung der Philosophie entgegenzuhalten, welche unsere Natur als Wesen der Freiheit und Autonomie über unser divergierendes Fürwahrhalten und unserer diskursiven Vernunftnatur und damit auch und insbesondere dadurch zu erfassen sucht, dass wir konstitutiv soziale Wesen sind (Brandom, 2019; Gabriel, 2020a; Meyer, 2024).[28] Demnach sind Gründe keine a-sozialen Entitäten die von einem Subjekt gewissermaßen logisch privat aufgenommen und angeeignet werden. Im Gegenteil sind Gründe sozial vermittelte Entitäten und der Kontext ihrer Vermittlung ist der Diskurs. In diesem Sinne könne es prinzipiell keine a-soziale, epistemologisch private Identifikation mit den Gründen unseres Denkens und Handelns geben. Stattdessen ist unser mentaler Innenraum, unsere Bedürfnisse, Gedanken, Konzepte und Wünsche bereits durch unsere Sozialisation sozial vermittelt, strukturiert und überformt (Gabriel, 2014, §8), (Gabriel, 2016a, S. 295ff.).[29] In Bezug auf das Verstehen haben Gründe als konstitutive Komponente der Person zwei epistemische Dimensionen. Einmal gibt es die subjektive Auseinandersetzung und ggf. Identifikation mit diesen Gründen, andererseits gibt es die epistemisch objektive (das heißt mitteilbare und dritt-personal erfassbare) Seite der Gründe.[30] Gründe sind mindestens partiell auch *öffentlich*. Andernfalls könnten sie gar nicht bestimmend sein, zum Beispiel für die Begründung von Gerichtsurteilen.[31] In der Optik dieser Anthropologie sind Menschen als die bisher *einzigen* Adressaten von Gründen, die Instanz, der die Verantwortung zum Herstellen eines Zustandes des Ver-

[28] „Subjekt-, Vernunft- und Freiheitsbegriff haben ihren primären Ort nach meinem Eindruck vielmehr in vernünftigen Diskursen zwischen Subjekten, die sich als gleichberechtigte Partner im Ringen um die besseren Argumente wahrnehmen – nicht als komplexe Objekte, deren Verhalten es vorherzusagen und zu erklären gilt. Subjekte und ihre Handlungen sind damit zunächst kategorial ganz anders bestimmt als als Gegenstände empirischer Theorien.“ (Meyer, 2024, S. 30) Für eine ontogenetische Ausarbeitung der These, dass unsere moralischen und kognitiven Fähigkeiten konstitutiv einen sozialen Anteil haben siehe (Tomasello, 2020).

[29] Damit möchte ich aber keinesfalls suggerieren, dass das Individuum auf die Gesellschaft zu reduzieren ist (siehe auch (8.2.5)). Diese Verhältnisse genauer auszubuchstabieren ist hier jedoch nicht der Platz.

[30] Ich habe mich bei dieser Distinktion inspiriert von der aus der Philosophie des Geistes stammenden Unterscheidung zwischen Bewusstseinszuständen die ontologisch subjektiv und zugleich epistemisch objektiv sein können (Nagel, 1978, 1986; Searle, 1989).

[31] Begrifflich liegt dies im Begriff Person angelegt, von Persona, das *was man sehen kann*. Unser Personsein hat folglich eine konstitutiv öffentliche Komponente. „Der Mensch ist danach eine mit der Fähigkeit zu eigenverantwortlicher Lebensgestaltung begabte 'Persönlichkeit'. Sein Verhalten und sein Denken können daher durch seine Klassenlage nicht eindeutig determiniert sein. Er wird vielmehr als fähig angesehen, und es wird ihm demgemäß abgefordert, seine Interessen und Ideen mit denen der anderen abzugleichen. Um seiner Würde willen muß ihm eine möglichst weitgehende Entfaltung seiner Persönlichkeit gesichert werden.“ (Becker, 1996, S. 35)

stehens zukommt.[32] Daraus leite ich ab, dass Gegenseitigkeit und Verantwortlichkeit als relationaler Prozess wesentlich für den Zustand des Verstehens ist. Hierfür müssen wir zwischen erklärungsgebender Instanz (in der Regel der Betreibende und das KI-System) und erklärungsempfangender Instanz (dem betroffenen moralischen Subjekt) in einem fraglichen Fall unterscheiden (BSI, 2022).[33] Die verantwortlichen Menschen sind in der Rechenschaftspflicht und damit fällt ihnen die Aufgabe zu, geeignete Verfahren zur Herstellung dieses Zustandes zu identifizieren und anzuwenden. Damit das betroffene Subjekt von dem Zustand der Fremdbestimmung in den Zustand der Selbstbestimmung kommen kann, muss eine Rechenschaftspflicht von Seiten der erklärenden Instanz in Form einer *Pflicht zum Erklären und Antworten* etabliert sein. Das entscheidende Moment dabei ist allerdings, dass die erklärungsgebende Instanz nicht nur einseitig Bestimmungen festlegen kann, wann ein Zustand des Verstehens erreicht ist. Das betroffene Subjekt muss diesen Prozess auch anerkennen. Interpretierbarkeit und andere Güter wie Sicherheit und Transparenz werden damit zu einer reziproken und dialogischen Angelegenheit. „Responsibility is not only about doing something and knowing what you're doing; it also means answerability. It is also a relational and communicative, perhaps even dialogical matter. The responsibility patient is the addressee, the one who is addressed in the responsibility relation." (Coeckelbergh, 2019) Die Maschine ist ein relevanter Teil in diesem Geflecht, da etwa ihre Robustheit und Nachvollziehbarkeit darüber entscheidet, wie gut die Verantwortlichen ihrer Pflicht gerecht werden können, doch sie automatisieren die menschliche Verantwortung nicht. Dort, wo Maschinen diese Aufgaben vollautomatisiert übernehmen, können wir von einem moralischen Defizit sprechen, da dies die Auflagen eines autonomiebasierten Menschenbildes verletzt (European Union (EU), 2016, (71)).[34] In diesem Sinne ist es rechtsethisch nicht vertretbar, wenn es keine verantwortliche natürliche Person gibt, die der Rechenschaftspflicht nachkommt.

[32] Dabei würde ich ihm Rahmen der unten noch in Ansätzen skizzierten Ontologie vorläufig dafür votieren, dass unsere freie, diskursive Vernunftnatur evolviert, indem wir als endliche epistemische Wesen das Problem lösen mussten, irreduzibel ungleiche Perspektiven verschiedener Individuen in Interaktion miteinander und zugleich mit der hyperkomplexen Wirklichkeit zu koordinieren.

[33] Dies wird vom Gesetzgeber auch durchaus so vorgesehen (European Union (EU), 2024).

[34] Dies ist auch mit der Gesetzgebung konsistent, da eine vollautomatisierte Entscheidung, die die Zuteilung von grundrechtsrelevanten Ressourcen betrifft, nach Erwägungsgrund (71) DSGVO auch rechtlich nicht zulässig ist.

4.6 Autonomie und Vertrauen

Die Diskussion des Instrumentalisierungsverbots oben hat bereits auf eine zentrale Problematik im Verhältnis von Transparenz und Autonomie hingedeutet. Die enge begriffliche Verknüpfung von Autonomie und Instrumentalisierungsverbot impliziert, dass eine betroffene Person, die Zwecke und damit verbundenen Daten und Prozesse auch verstehen kann. Im Kontext der künstlichen Intelligenz beispielsweise kann dies bedeuten, dass ein betroffenes Subjekt sich über die statistischen Ableitungen eines Kreditvergabesystems im Klaren ist und zustimmt, ob sie von diesem System bewertet werden möchte oder nicht. Sie versteht bis zu einem gewissen Grade die Gesetzmäßigkeiten, die indirekt das eigene Leben, den Zugang zu Ressourcen regeln und stimmt diesen Gesetzmäßigkeiten, im Wissen der bestehenden Risiken und Ungenauigkeiten, zu. Dieser Anspruch drückt sich in der vieldiskutierten Formulierung in der DSGVO aus, dass Menschen, die von automatisierten Entscheidungen betroffen sind, ein Recht haben, Kenntnisse über die Logik hinter dem Entscheidungsprozess zu bekommen, als auch über die Tragweite und Konsequenzen solcher Prozesse informiert zu werden.[35] Begrifflich umfasst *Selbst-Gesetzgebung* das Verständnis dieser Gesetze (oder „involvierten Logiken“ nach DSGVO) bzw. wirft die Frage auf, wie eine autonome Einwilligung in Prinzipien aussehen kann, die die betroffene Person nicht versteht. Dies mutet zunächst kontradiktorisch an. Folgt nicht aus der Intransparenz bestimmter Prozesse, dass ein von diesen Prozessen betroffenes Subjekt notwendig instrumentalisiert wird und letztlich nicht Selbstbestimmung (Autonomie), sondern Fremdbestimmung (Heteronomie) vorliegt. Dies ist letztlich der Grund, warum ich vorschlagen möchte, dass die unterschiedlichen rechtsethischen Güter an der Schnittstelle aus Sicherheit und Interpretierbarkeit sich zunächst im Anspruch des Verstehens ausdrücken, welcher sich unter realen Bedingungen letztlich im normativen Anspruch nach Vertrauen transformiert (siehe auch (5.9)). Die Wendung *vertrauenswürdige KI* stellt

[35] „Die betroffene Person sollte das Recht haben, keiner Entscheidung — was eine Maßnahme einschließen kann — zur Bewertung von sie betreffenden persönlichen Aspekten unterworfen zu werden, die ausschließlich auf einer automatisierten Verarbeitung beruht und die rechtliche Wirkung für die betroffene Person entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt, wie die automatische Ablehnung eines Online-Kreditantrags oder Online-Einstellungsverfahren ohne jegliches menschliche Eingreifen. Zu einer derartigen Verarbeitung zählt auch das „Profiling“, das in jeglicher Form automatisierter Verarbeitung personenbezogener Daten unter Bewertung der persönlichen Aspekte in Bezug auf eine natürliche Person besteht, insbesondere zur Analyse oder Prognose von Aspekten bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben oder Interessen, Zuverlässigkeit oder Verhalten, Aufenthaltsort oder Ortswechsel der betroffenen Person, soweit dies rechtliche Wirkung für die betroffene Person entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt.“ (European Union (EU), 2016, (71))

auch eine Konstante in der KI-Verordnung dar (European Union (EU), 2024). Vertrauen ist das komplexe Moment, wenn eine Person zunächst begrenzt Kenntnisse über die sie betreffenden Prozesse erlangt, doch im Bewusstsein der partiellen Opazität dieser Prozesse in sie einwilligt (Brandom, 2019; Luhmann, 1968). Genau diese Konstellation aus der Anerkennung des Wissens *und* dem Bewusstsein des Nicht-Wissens ist Vertrauen und belebt damit den Raum zwischen Autonomie und Heteronomie. Dem Moment des Vertrauens in Interaktion mit hochkomplexen Technologien und Umwelten werden wir in den kommenden Kapiteln noch weiter nachspüren. Auf die Technologie angewendet, sind zunächst zwei Bedingungen als notwendig (aber keinesfalls hinreichend siehe (8.1.1)) für den Zustand des Vertrauens festzuhalten: *a.)* die betreffende Person hat ein hinreichendes Bewusstsein über die sie betreffenden Gesetzmäßigkeiten (Algorithmen, Entscheidungslogiken etc.) und *b.)* die Person hat ein hinreichendes Bewusstsein über die Opazität und die mit dieser einhergehenden Risiken. Dies wirft unmittelbar die Frage auf, was ein *hinreichendes* Bewusstsein von *a.)* den entsprechenden Gesetzmäßigkeiten und *b.)* der Intransparenz bzw. Opazität dieser genauer charakterisiert. Mit anderen Worten: Was genau bedeutet es, zu verstehen – und was bedeutet es, zu vertrauen?

4.7 Informationelle Selbstbestimmung und Integrität

Wenn wir das Gesagte auf einen Begriff bringen wollen, würde ich vorschlagen, dass es sich um eine Interpretation der Ausdrücke Informationelle Selbstbestimmung und Integrität handelt. Die Selbstbestimmung und Integrität ist ganz elementar mit dem Instrumentalisierungsverbot, der Menschenwürde und der Autonomie assoziiert (*Beschluss des Bundesverfassungsgerichts: "Volkszählungsurteil" (Recht auf informationelle Selbstbestimmung)*, 1983). Insbesondere im Datenschutz drückt sich aus, dass im Zeitalter der Digitalisierung die persönliche Integrität stark mit der informationellen assoziiert ist (Caspar, 2023, S. 239ff.). Die Freiheit und Autonomie hängt elementar von der Integrität der Daten ab. Menschen sollen autonom über die Art und Weise, wie Informationen und deren Verarbeitung ihr Leben bestimmen, entscheiden können. Sie sollen damit ihre Autonomie und personale Integrität über alle relevanten Kontexte behalten können. Dies ist ein ethisch wie auch praktisch äußerst hoher Anspruch, der sich aber unmittelbar aus den zentralen Prämissen eines Menschenbildes der Mündigkeit und Würde ableitet.

5 Interpretierbarkeit als epistemische Sicherheit

Aus Gründen, die an anderer Stelle erläutert sind, ist der Ausdruck Interpretierbarkeit nicht rigoros mathematisch auflösbar und dies soll hier auch nicht verfolgt werden (Z. C. Lipton, 2016) (und siehe auch (5.7)). Ich möchte hier aber dennoch eine theoretisch fundierte und partiell formalisierbare Begriffsanalyse von Interpretierbarkeit anbieten: die *pluralistische Bedingungsontologie*, die sich in den Sprachen der Mengenlehre, Logik und Wahrscheinlichkeitstheorie anteilig formalisieren (5.4) und bis zu einem gewissen Grad quantitativ übersetzen und testen lässt (7.5) (Ellis, 2012; Gabriel, 2016c; Mothilal & Tan, 2021).

5.1 Sicherheit und Verstehen als epistemischer Zustand

Der Begriff Interpretierbarkeit überlappt semantisch mit einer Vielzahl anderer Ausdrücke, wie Verstehen, Überzeugung, Erklärung, Rechtfertigung und viele mehr. Es ist mittlerweile viel über die Ambiguität und die semantische Vernetzung dieser Ausdrücke geschrieben worden und die Argumente sollen hier nicht nochmal im einzelnen wiederholt werden (P. Lipton, 2004; Z. C. Lipton, 2016; Miller, 2019). Stattdessen schlage ich eine pragmatische Lösung für dieses Problem vor. Es geht mir hier nicht darum, die linguistischen Facetten eines Wortes zu thematisieren, sondern darum, einen epistemischen Zustand zu charakterisieren, nämlich der, in dem sich ein Subjekt im Moment des Verstehens befindet. Das heißt, wenn ein Subjekt S ein Ereignis Y (Kreditvergabe, Sicherheitsanalyse etc.) erfolgreich interpretiert (bzw. versteht), welche Bedingungen sind dann erfüllt, dass S sich im epistemischen Zustand des *Verstehens* befindet, bzw. wie können wir diesen Zustand charakterisieren. Und als Titel für diesen Zustand nutze ich hier den Ausdruck Verstehen oder manchmal auch Interpretieren. Wir werden allerdings sehen, dass sich aus dieser Analyse durchaus Differenzierungsmerkmale ergeben, zum Beispiel zwischen dem Zustands des Verstehens und dem des Wissens, Rechtfertigens oder Erklärens. Mit dieser methodischen Weichenstellung können wir Probleme, die aus der Schnittmenge von Sicherheit und Interpretierbarkeit resultieren präzise als epistemische Problemstellungen fassen, in der ein epistemisches Subjekt (bzw. Subjekte) in direkter oder (in)direkter Interaktion mit Maschinen in den Zustand des (nicht-)Verstehens gerät, während das gleiche Subjekt gerechtfertigte (moralische, rechtliche usf.) Anforderungen an das Verstehen

stellen kann. Wenn zum Beispiel eine Bewerbung auf eine Arbeitsstelle automatisiert durch ein KI-System abgelehnt wird und der betroffenen Person aufgrund der Blackbox-Natur des Systems keine befriedigende Erklärung für diese Entscheidung erhalten kann, dann befindet sich die Person (das epistemische Subjekt) in einem Zustand des Nicht-Verstehens bzw. des Nicht-Verstehen-Könnens, während es gleichzeitig moralische und auch rechtliche (DSGVO) Anforderungen an ein solches Urteil stellen kann (Goodman & Flaxman, 2016).

5.2 Der Satz vom zureichenden Grunde

Zum Aufbau eines analytischen Modells von Interpretierbarkeit möchte ich hier eine Auffassung vertreten, die als eine zeitgenössische Interpretation von Leibniz' Satz vom zureichenden Grunde verstanden werden kann (Gabriel, 2016a; Leibniz, 1998).

Postulat 1: Der Satz vom zureichendem Grunde besagt in seiner allgemeinsten Form, dass wenn etwas geschieht, „dass keine Tatsache als wahr oder existierend gelten kann und keine Aussage als richtig, ohne dass es einen zureichenden Grund dafür gibt, dass es so und nicht anders ist, obwohl uns diese Gründe meistens nicht bekannt sein mögen." (Leibniz, 1998, §32)

Das heißt wiederum, wenn ein Ereignis Y (Kreditvergabe, Sicherheitsanalyse, Verfassen einer Abschlussarbeit etc.) geschieht, dann gibt es eine Reihe von Gründen (später Bedingungen), die erklären, dass es geschieht. Dieses Postulat unterstellt schlicht die Intelligibilität der Welt (Nagel, 2016, S. 30).[36]

5.3 Bedingungsontologie

Der Vorzug dieser Annahme ist, dass der Satz vom zureichenden Grunde noch nicht auf eine bestimmte Ontologie restringiert ist (siehe insbesondere (6.1)), das heißt die *Gründe* für ein *Ereignis* sind nicht notwendigerweise ausschließlich natürliche (physikalisch, chemische, biologische), psychologische (motivational, emotional, behavioristisch und so weiter), soziologische (Schwarmverhalten, doppelte Kontingenz und so weiter) oder andere Gründe. Vielmehr können wir mit dem Satz vom zureichenden Grunde problematische Reduktionismen (Ellis, 2016) umgehen und die Pluralität von *Gründen* für Ereignisse

[36]Der Nachweis, dass diese Annahme selbst gegen harte Spielarten des Skeptizismus, das heißt der prinzipiellen Möglichkeit der Unmöglichkeit von Wissen gelingt, ist an anderer Stelle gezeigt worden. Für eine analytisch bemerkenswerte Verteidigung dieser These siehe (Gabriel, 2016a, S. 352ff.).

in einem holistischen Bottom-up und Top-down Verhältnis anerkennen (Gabriel, 2020a; Voosholz & Gabriel, 2021) (6.1). Erklären können wir vorläufig erstmals als das begriffliche Erfassen solcher Gründe definieren, die dann als Prämissen Elemente von juristischen, wissenschaftlichen, technischen und sonstigen Begründungen werden können.[37]

Die Verwendung des Ausdrucks Grund deutet bereits an, dass es sich hierbei um intelligible Bedingungen handelt, folglich, dass diese von den meisten betroffenen Menschen unter geeigneten Bedingungen verstanden und in vernünftigen Diskursen explizit gemacht werden können. Zum Beispiel ist die in ganzen, positiven Zahlen angegebene Höhe des Einkommens eine Variable, die von den meisten Kreditbewerbenden verstanden werden kann (was natürlich nicht impliziert, dass sie auch die Rolle, die diese Variable in einem Vergabeverfahren spielt, akzeptieren). Der Ausdruck Grund suggeriert auch eine Engführung auf sprachliche Diskurse. Stattdessen bietet es sich an, anstelle von Gründen von Bedingungen zu sprechen. Der Ausdruck Bedingung ist im informationswissenschaftlichen Kontext aus folgendem Grund noch einmal von besonderem Vorteil. Wie wir in den folgenden Kapiteln noch sehen werden (6.3), sind nicht alle Bedingungen in diesem Sinne intelligible, was für datenverarbeitende Systeme noch einmal eine besondere Rolle spielt. Der Bedingungsbegriff erlaubt vielmehr recht unkompliziert alle möglichen bedingenden Faktoren zu thematisieren. Präzisieren wir diese Überlegung als *Postulat 2*:

Postulat 2 Bedingungsontologie: Wenn etwas (Handlungen, Ereignisse, Prozesse, Gegenstände und so weiter) existiert, dann gibt es *a.)* notwendige Bedingungen für dessen Existenz, die *b.)* zusammengenommen hinreichend sind, damit es existiert (Gabriel, 2015, S. 285ff.), (Gabriel, 2020a, § 13).

Im Folgenden spreche ich immer, wenn wir ein zu erklärendes Explanandum Y haben von einem Ereignis, Prognose oder Output. Dies ist in dieser Arbeit nur der Name für etwas, das existiert, wofür wir potenziell einen Anspruch auf Interpretierbarkeit stellen. Es kann sich also um eine Kreditprognose, eine biometrische Identifizierung, das Ergebnis einer Sicherheitsanalyse, eine Bundestagswahl oder das Ergebnis einer Differentialgleichung handeln. Alles, was wir potenziell erklären wollen, sei als Ereignis Y verstanden.

[37] Aus Platzgründen spare ich die hier noch deutlich komplexere Frage der Willensfreiheit aus, die dieser Analyse noch einen weiteren Layer der Komplexität hinzufügen würde (siehe auch (8.2.5)). Doch da diese Arbeit von der Interpretierbarkeit von technischen Artefakten handelt und nicht der Interpretierbarkeit menschlicher Handlungen, die wahrscheinlich etwas fundamental anderes sind, kann dies hier ausgespart werden.

5.4 Formale Bedingungen für Interpretierbarkeit

Präzisieren wir nun das *Postulat 2* formal.[38]

Postulat 2: Wenn Y existiert, dann gibt es notwendige Bedingungen $X_1, X_2, \ldots X_n$, die zusammengenommen hinreichend sind, damit es existiert.

Präzisieren wir weitergehend notwendig und hinreichend:

1. $X_1 \ldots X_n$ ist eine notwendige Bedingung für Y, wenn gilt:

$$Y \Rightarrow X$$

   In Worten: Y kann nicht der Fall sein, wenn X nicht der Fall ist. Wenn X der Fall ist, muss Y aber nicht zwingend der Fall sein.

2. $X_1 \ldots X_n$ ist eine hinreichende Bedingung für Y ist wenn gilt:

$$X \Rightarrow Y$$

   In Worten: Wenn X der Fall ist, dann muss Y zwingend der Fall sein. X alleine ist hinreichend für Y.

Daraus folgt, dass Notwendigkeit und Suffizienz logisch kontrapositiv zueinander sind. Nach dem Gesetz der Kontraposition können beide Definitionen umformuliert werden:

$$X \Rightarrow Y \quad \text{äquivalent zu} \quad \neg Y \Rightarrow \neg X$$

$$Y \Rightarrow X \quad \text{äquivalent zu} \quad \neg X \Rightarrow \neg Y$$

In Worten: X ist hinreichend, wenn es nicht einen Fall gibt, da Y nicht der Fall ist und X auch nicht der Fall ist. X ist notwendig, wenn X nicht der Fall ist, dann ist Y auch niemals der Fall.

[38] Die Formalisierungen sind inspiriert und orientiert an (Halpern, 2016; Mothilal & Tan, 2021).

5.5 Kausalität und Interpretierbarkeit

Wenn Subjekte gerechtfertigte Anforderungen an Interpretierbarkeit innerhalb der (in)direkten Relation zu Maschinen stellen, dann kommt Kausalität in dieser Beziehung eine entscheidende Rolle zu. Begründen wir dies noch einmal am obigen Beispiel Kreditvergabe.

Narrative Plausibilität und Erklärungspluralismus Beispiel Kreditvergabe: Nehmen wir an, ein Antragstellender bekommt nach Abschluss des Prüfverfahrens keinen Kredit ($\neg Y$). Wenn dabei Datenanalyse und/oder maschinelles Lernen zum Einsatz kam, dann hat der Kreditbewerbende laut Gesetzgebung in der europäischen Union unter Umständen das Recht eine Erklärung zu erhalten (Goodman & Flaxman, 2016). Nun stellt sich die Frage, was hier *als* Erklärung zählt.

Eine zentrale Herausforderung bei der Erklärung solcher Prozesse ist das, was wir unter dem Begriff narrative (Schein-)Kausalität/Plausibilität oder retrospektiven Erklärungspluralismus fassen können (P. Lipton, 2004; Salmon, 1984). Dies ist der Umstand, dass:

a) für ein und dasselbe Y retrospektiv eine Pluralität von (mindestens scheinbar) konsistenten Erklärungen $X_1 \dots X_n$ anbieten lassen, die plausibel beide Bedingungen unter (5.4) erfüllen könnten.

b) für verschiedene Outputs $Y_1 \dots Y_n$ sich die gleiche Erklärung anbieten lässt.

Fall 1: Retrospektive Erklärungspluralität für dasselbe Ereignis Y Eine Antragstellerin erhält keinen Kredit ($\neg Y$). Mögliche Erklärungen ($X_1 \dots X_n$), die alle formal als hinreichend (oder notwendig) gelten könnten:[39]

- X_1: Die Antragstellerin verfügt über kaum finanzielle Rücklagen (Kontostatus)
- X_2: Die Antragstellerin besitzt keine Immobilie (Vermögen)
- X_3: Die Antragstellerin hat weitere laufende Kredite (Kreditgeschichte)

Fall 2: Gleiche Erklärung für unterschiedliche Outputs $Y_1 \dots Y_n$ In diesem Fall sind die Eingangsvariablen ($X_1 \dots X_n$) bei verschiedenen Kreditbewerbenden identisch, aber der Output variiert (im Beispiel $Y/\neg Y$). Bei linearen Modellen liegt dies darin begründet, dass

[39] Ich nutze der Einfachheithalber für dieses Beispiel die weibliche Form.

die einzelnen Variablen eine unterschiedliche Gewichtung erhalten. So könnte zum Beispiel bei einem Kreditinstitut die Variable *regelmäßiges Einkommen* im Verhältnis zur Höhe des Einkommens stärker gewichtet werden, als bei einem anderen Institut.

In beiden Fällen kann eine Erklärung unter Umständen moralisch relevant und/oder sogar rechtlich geboten sein. Die beiden Fälle zeigen auf, dass eine *retrospektive* konsistente Erklärung in notwendige und hinreichende Bedingungen alleine wiederum selbst nicht hinreichend ist, um als Erklärung zu gelten. Es ist besonders wichtig, Erklärbarkeit/Interpretierbarkeit und Rechtfertigung zu differenzieren. Das retrospektive, sachlich fundierte Rechtfertigen eines Ereignisses ist in diesem Modell noch kein Verstehen. Nichtsdestotrotz ist ein hoher Grad an sachlogischer narrativer Plausibilität wahrscheinlich notwendig für erfolgreiches Verstehen (siehe auch (7.2.5)).

Kausalität und Autonomie Wie dargestellt (4.2), ist Autonomie verbunden mit einem instrumentellen Technikverständnis konstitutiv für die rechtsethische Technologiekonzeption. Dabei sind in der Schnittmenge aus Autonomie und Erklärbarkeit zwei Dimensionen ihrer Temporalität nach relevant. Einerseits geht es darum, in der Retrospektive (*ex post*) eine konsistente Erklärung für ein Ereignis anzubieten. Anderseits geht es darum, prospektiv (*ex ante*) über die bedingenden Faktoren für ein Ereignis aufgeklärt zu sein.[40] Beide Perspektiven können als wesentlich für ein autonomes Verhältnis zu Technologie gelten. Letzteres erlaubt auf die Zukunft gerichtetes autonomes Handeln, ersteres erlaubt rückschauendes Verstehen und damit eine retroreflektive Emanzipation. Es lässt sich argumentieren, dass die freie Entfaltung der Persönlichkeit in modernen Gesellschaften unter Bedingungen der Digitalisierung von beiden Dimensionen abhängt. Doch zeigt die Retrospektive ein entscheidendes Defizit in Bezug auf Autonomie. Die *Erklärungspluralität* hinterlässt die betroffenen Subjekte zunächst in der passiven Situation einer narrativen Plausibilität, doch eben auch nicht mehr. Damit die retrospektive Erklärung autonomie-fördernd wirkt, lässt sie sich dahingehend stärken, dass sie um eine konditionale Komponente ergänzt werden sollte. Die Frage ist nicht nur, wie ein bestimmter Output zustande kam, sondern unter welchen Bedingungen sich ein gegebener Output verändert hätte. In anderen Worten, wie hätte ein betroffenes Subjekt den *kausalen Verlauf* der Dinge verändern können (Halpern, 2016; Mothilal & Tan, 2021).

[40] Diese Unterscheidung ist inspiriert von (Coeckelbergh, 2019; Dignum, 2019).

Formalisierung kontrafaktische Kausalität Hier können wir nun das Kausalitätskonzept aus der Komplexitätsforschung, Physik und Informatik mobilisieren (siehe auch (6.1)):

„Causes are separated from effects by searching for correlations between phenomena such that manipulation of one (the cause) can be shown, in a specific context, to reliably result in specific changes in the other (the effect) at a later time. One has to search for this correlation in the midst of internal and environmental noise. Laboratory tests of isolated systems allow an understanding of the elements of causation, which are interactions between the particles that underlie all physical existence.“ (Ellis, 2016, S. 8)

„Existence. We must recognise the existence of any kind of entity that demonstrably has a causal influence on physical systems.“ (Ellis, 2016, S. 14)

„The relation of causation that can hold between entities consists in the fact that variation of properties in one object systematically changes the properties in the other object.“ (Voosholz & Gabriel, 2021, S. 8)

Im formalen Vokabular unseres Modells lässt sich dies wie folgt formulieren:

1. Korrelation zwischen X und Y (statistische Voraussetzung)

$$\text{Corr}(X, Y) \neq 0$$

Es besteht eine statistisch signifikante Korrelation zwischen X (Bedingungen) und Y (Output/Ereignis). Dies ist selbst eine notwendige, aber keine hinreichende Bedingung für Kausalität.

2. Kontrafaktische Kausalität

X_i ist eine kontrafaktisch-kausale Bedingung für Y, wenn gilt: $\neg Xi \rightarrow \neg Y$ (Kontrafaktum: Wenn X_i nicht vorläge, dann wäre Y nicht eingetreten) oder in kontrapositiver Form: $Y \rightarrow X_i$ (Notwendigkeit unter kontrafaktischer Perspektive).

Dabei ist die kontrafaktische Relation $\neg X_i \Rightarrow \neg Y$ als epistemisch interpretierbar im Sinne einer möglichen Manipulierbarkeit von X_i zur Erreichung eines gewünschten Ereignisses

Y. Formal ergibt sich daraus die Definition kontrafaktisch-kausaler Relevanz: Eine Bedingung X_i ist kontrafaktisch-kausal relevant für Y genau dann, wenn gilt: Kontrafaktische Abhängigkeit: $Y \wedge \neg X_i$ ist logisch inkonsistent.

Erläutern wir dies direkt an einem Beispiel. Wir stellen nun kontrafaktisch die Frage: Wäre Y (Kreditbewilligung) erfolgt, wenn X_1 (zum Beispiel ein regelmäßiges Einkommen) gegeben gewesen wäre? Wenn die Antwort „ja“ lautet, dann ist X_1 als kontrafaktisch-kausale Bedingung relevant für Y (Dandl et al., 2020). Dieses Beispiel illustriert die in gewissem Sinne selbstermächtigende Wirkung einer solchen Erklärung. Sie erlaubt es einem betroffenen Subjekt die Frage sinnvoll zu beantworten: 'Welche Bedingungen hätten vorliegen müssen, um ein anderes Ereignis (zum Beispiel eine positive Kreditprognose) zu erreichen?'

Bedingungsontologie und Kausalität Diese Besprechung zeigt auf, dass sich die Bedingungsontologie als Kandidat für ein Modell der Kausalität qualifiziert, die im Einklang mit dem oben angeführten physikalischen und philosophischen Verständnis von Kausalität ist. Sie erlaubt es, Bedingungen als kausale Faktoren zu identifizieren, indem sie als Möglichkeitsrelation in einem Modellraum von Bedingungen $X_1, \ldots X_n$ verstanden werden. Kontrafaktische Kausalität ergänzt diese Perspektive auf Bedingungen um eine Dimension der epistemischen Güte: Bedingungen werden hinsichtlich ihrer Erklärungskraft bewertet, also in Bezug darauf, wie sie in kontrafaktischen Szenarien kausal für ein Ereignis Y verantwortlich gemacht werden können.

5.6 Mechanistische Erklärbarkeit

Wenn wir den Ausführungen bis hierhin glauben schenken, dann etabliert sich konsequent ein idealisiertes Szenario von Erklärbarkeit, welches ich hier algorithmisch im informatischen Sinne nennen möchte (in der englischen Literatur oftmals auch mechanistisch genannt, von *mechanistic interpretability*). „Ein Algorithmus ist eine *endliche*, *eindeutig* definierte Folge von Anweisungen oder Rechenschritten zur Lösung eines Problems oder zur Durchführung einer Aufgabe (meine Hervorhebungen, J.N.).“ (H. Müller & Weichert, 2023, S. 16f.)[41] Übersetzt unter Rückgriff auf *Postulat 2* wäre eine Erklärung dann und nur dann

[41] Ich gehe der Einfachheit halber von einem idealisiert-deterministischen Konzept von Algorithmen aus und klammere hier bewusst stochastische und probabilistische Algorithmen aus (Ernst, Schmidt & Beneken, 2023, S. 555f.).

algorithmisch, wenn die angegebenen Bedingungen $X_1, X_2, \ldots, X_n$ den zu erklärenden Prozess Y als notwendige und hinreichende determinieren. Daraus folgt nach dem Prinzip der Kontraposition, wenn Y der Fall ist, dann wissen wir, dass auch $X_1, X_2, \ldots, X_n$ der Fall ist und wenn $X_1, X_2, \ldots, X_n$ vorliegen, dann gilt auch Y.

Beispiel: Kreditbewilligung durch ein KI-System Nehmen wir an, ein KI-System entscheidet automatisch über die Bewilligung eines Kredits (Y). Wir verwenden wieder Variablen aus dem Evaluations-Datensatz (7.5.6). Die Bedingungen könnten sein:

- X_1: Kreditgeschichte: Der Antragsteller hat keine anderen laufenden Kredite.
- X_2: Sparkonto: Der Antragsteller hat ausreichend Rücklagen auf einem Sparkonto gebildet.
- X_3: Vermögen: Der Antragsteller verfügt über Sicherheiten in Form einer Immobilie.

Notwendigkeit: Wenn der Kredit (Y) gewährt wird, müssen diese Bedingungen (X_1 bis X_3) vorliegen. Fehlt eine dieser Bedingungen, ist die Entscheidung zur Kreditvergabe nicht gerechtfertigt: $\neg X \Rightarrow \neg Y$.

Suffizienz: Sind alle Bedingungen gegeben, folgt daraus zwingend die Kreditbewilligung ($X \Rightarrow Y$). Die Entscheidungslogik des KI-Systems wäre in diesem idealisierten Szenario *mechanistisch* erklärbar.

5.7 Komplexität und die Grenzen der Erklärbarkeit

5.7.1 Mechanistische Erklärbarkeit versus vollständige Interpretierbarkeit

Als wichtige Einschränkung ist hier bereits darauf hinzuweisen, dass eine *mechanistische* Erklärung nicht äquivalent ist mit einer *vollständigen* Interpretierbarkeit (für mehr Details siehe (6.3)). In dem Beispiel oben würde aus *endlich vielen Fällen* beobachteten Modellverhaltens eine *externe Entscheidungslogik* formal rekonstruiert werden. Diese besagt zum Beispiel, wenn kein regelmäßiges Einkommen vorliegt, dann wird der Kredit nicht bewilligt. Die Semantik dieser Entscheidungslogik ist für uns verständlich, bedeutet aber nicht, dass sie *isomorph* zur *internen* Entscheidungslogik des Modells ist. Machen wir dies an einem Beispiel deutlich. So könnte das Modell anhand der Trainingsdaten aufgrund einer

systematischen Scheinkorrelation zwischen Einkommen und Postleitzahl gelernt haben, dass Personen mit Postleitzahl (PLZ) im Bereich `70025` meist ein hohes Einkommen besitzen. Das Modell hätte die Postleitzahl somit als stellvertretendes Merkmal (*Proxy Variable*) für Einkommen gelernt (Tu et al., 2020; Ye et al., 2024).

Nun haben wir für unsere Testung zur Modellevaluation zufällig nur Profile mit PLZ 70025 genommen.

Aus der rekonstruierten Entscheidungslogik stellt sich diese für uns extern wie folgt dar:

$$\neg X_1 \Rightarrow \neg Y$$

Bzw.: Wenn kein ausreichendes Einkommen vorliegt, dann wird der Kredit nicht bewilligt.

Obgleich die *interne* Entscheidungslogik eigentlich lautet:

$$\neg \text{PLZ } 70025 \Rightarrow \neg Y$$

Wenn wir das Modell nach einer solchen Evaluation für den praktischen Einsatz zulassen, könnten Personen aus anderen Stadtteilen, trotz ausreichendem, regelmäßigen Einkommens, keine Kreditbewilligung erhalten. Damit wäre die Semantik der externen mechanistischen Entscheidungslogik falsifiziert.

5.7.2 (Hyper-)Komplexität

Die aus den technischen Eigenschaften der Modelle resultierenden Probleme werden uns weiter unten noch beschäftigen (6.3). Darüber hinaus gilt es kurz auf die grundsätzlichen Limitationen eines mechanistischen Verständnisses von Interpretierbarkeit zu sprechen zu kommen. Wenn wir im obigen Sinne alle Bedingungen $X_1, \ldots, X_n$, die die Wahrheit einer Erklärung für ein Ereignis Y garantieren, kennen, dann wären dies wahrheitsgarantierende Gründe. Nun hat die Erkenntnistheorie und Komplexitätsforschung hinreichend bewiesen, dass es solche absoluten Gründe prinzipiell nicht geben kann (Ellis, 2012; Gabriel, 2016a, u.a. S. 62f.). Die Einschränkungen für diese notwendige Endlichkeit der Gründe lassen sich

grob in zwei Klassen sortieren:[42]

Einschränkung 1: Komplexität Erstens sind die Bedingungen aufgrund fehlender Ressourcen (Zeit, Instrumente, Voreingenommenheit, tradierte wissenschaftliche Tugenden und vieles mehr) nicht vollständig einholbar. Auch unter der Annahme, dass diese prinzipiell zu ermitteln und sogar interpretierbar wären besteht immer die Möglichkeit, dass wir bestimmte Gründe noch nicht berücksichtigt haben, die damit im Prinzip das Potenzial bergen, eine angegebene Erklärung zu falsifizieren (Gabriel, 2016a, S. 95f.), (Popper, 1934). Zu dieser Klasse können wir auch eine weitere Einschränkung hinzuzählen, welche uns weiter unten noch interessieren wird (6.1). Selbst wenn alle Informationen einholbar wären, werden viele Prozesse in Natur, Gesellschaft, Psychologie und Informatik überhaupt erst dadurch verständlich, dass Informationen von einem System gekapselt und dessen Prozesse vor anderen Systemen verborgen bleibt. Viele dieser Prozesse, wie Computerprogramme, die Homöostase von Lebewesen oder das Planen von großen Bauprojekten können wir nur verstehen, da die konstitutiven Bedingungen des Systems von dem ausführenden Prozess verborgen sind, andernfalls würden diese Prozesse kollabieren (Ellis, 2016, S. 41f.). Die Erkenntnis des einen Prozesses wird erst möglich, durch die Unmöglichkeit der gleichzeitigen Erkenntnis seiner Bedingungen. Diese Klasse von Einschränkungen lässt sich treffend als Komplexität bezeichnen.

Einschränkung 2: Hyper-Komplexität Oder es gibt Gründe, die wir prinzipiell nicht transparent machen können, was wiederum an den Eigenschaften des zu untersuchenden Gegenstands selbst liegt oder an prinzipiellen Schranken unserer kognitiv-epistemologischen Konstitution.[43] Das heißt die Fallibilität einer Erklärung ist nicht durch verborgene Variablen oder endlichen Ressourcen bedingt, sondern folgt aus den Eigenschaften des Gegenstandes *an-sich*. Eine solche Eigenschaft ist beispielsweise das Phänomen echten Zufalls, das heißt die Auffassung, dass ein Ereignis keine weitere kausale Bedingung hat. Obzwar

[42] Dies ist eine höchst allgemeine Darstellung, die deutlich weiter differenziert werden könnte. Auch wird der mögliche Beitrag echter menschlicher Freiheit als Faktor innerhalb des Netzwerkes der Gründe als epistemische Schranke hier nicht weiter thematisiert.

[43] Letztere ist eine Argumentationslinie, welche in der Geschichte gewissermaßen ihren Höhepunkt in dem erkenntnistheoretischen Hauptwerk der Moderne, Kants *Kritik der reinen Vernunft* und der Diskussion um das *Ding an sich* findet. Dort finden wir eine systematische Untersuchung des menschlichen Verstandes als Bedingungen der Möglichkeit von Erkenntnis und auch was geschieht, wenn versucht wird, diese Bedingungen bzw. Restriktionen des Verstandes zu überschreiten (Kant, 1781). Zum Nachweis der Endlichkeit epistemischer Wesen wie dem Menschen siehe (Gabriel, 2014) und insbesondere auch (Hogrebe, 2006).

umstritten ist, ob es echten Zufall gibt, ist es unstrittig, dass unter den realen mesoskopischen Bedingungen unseres Lebens Ereignisse als mindestens hinreichend in-deterministisch einzustufen sind, dass sie vom echten Zufall für uns in keiner relevanten Hinsicht zu unterscheiden sind. Somit könnten wir prinzipiell über eine *stochastische Erklärungsgüte* nie hinauskommen. Eine zweite, in den letzten Jahren ausgearbeitete Argumentation richtet sich gegen die metaphysische Prämisse, dass die Wirklichkeit eine kausal-nomologisch geschlossene Totalität „die Welt“ (Gabriel, 2016c, S. 224ff.) ist. Wenn wir die Wirklichkeit (bzw. die Welt) als kausal-geschlossene Totalität verstehen, dann implizierte dieses Bild, dass die Bedingungen für ein Ereignis *ex hypothesi* dieses im Prinzip determinieren.[44] Genau genommen spräche wissenschaftlich und auch phänomenologisch viel mehr dafür, dass die Realität eben nicht kausal geschlossen, sondern konstitutiv offen ist. Geht der Monismus letztlich von einer zu vereinheitlichenden Endlichkeit der Gegenstände und Gegenstandsbereiche aus, so votiert der Pluralismus für eine infinite Proliferation der Gegenstände und Gegenstandsbereiche (Gabriel, 2016a, 2016c, 2020a; Schaffer, 2010, S. 276).[45] Im Einzugsbereich dieser Strategie liegt auch ein Argument, welches in Analogie zu den berühmten Unvollständigkeitstheoremen der Mathematik formuliert wurde. Dabei werden ihre Implikationen unter anderem dahingehend verallgemeinert, dass jedes theoretische Unterfangen von Bedingungen lebt, die es selbst nicht beweisen kann (Gabriel, 2014; Hogrebe, 2006; Kreis, 2015, Kap. II). So gilt auch für das Interpretieren von Ereignissen eine *prinzipielle* und nicht bloße *kontingente* Verstehensgrenze (für weitere Details siehe auch (8.2)). Diese Phänomene möchte ich hier mit dem Ausdruck Hyper-Komplexität fassen, worunter hier jedes Ereignis verstanden sei, welches prinzipiell nicht in einer deterministischen Erklärung aus notwendigen und hinreichenden Bedingungen aufgeht. In diesem Sinne sind hyper-komplexe Ereignisse in-deterministisch, das heißt nicht restlos in einer kausalen Analyse zu enthüllen.

[44] Wobei an dieser Stelle offengelassen werden kann, wie genau ein Determinismus in diesem Rahmen auszubuchstabieren wäre.

[45] Obgleich sich hier bei Gabriel womöglich eine Ambiguität auftut, die er selber in einem Vortrag auf Nachfrage konstatiert hat (Gabriel, 2016b, Min: 60). Einerseits meint er die prinzipiellen Offenheit der Gegenstandsbereiche (Sinnfelder in seiner Terminologie) nachgewiesen zu haben. Er nennt dies treffend auch die „Unvollständigkeit der Welt“ (Gabriel, 2016a, S. 370ff.). Andererseits vertritt er eine Auffassung des Satzes vom zureichendem Grunde, welcher identisch ist mit dem oben vorgestellten. Er fällt nicht notwendigerweise in einen Determinismus zurück, da er, ähnlich wie in dieser Arbeit vertreten, nicht alle Bedingungen für *harte Ursachen* hält.

5.8 Interpretierbarkeit als Komplexitätsreduktionsmechanismus

Wie besprochen, bietet die Wirklichkeit nur unter den rigoros idealisierten Bedingungen den Horizont des Verstehens im oben beschriebenen Sinne (5.6). Realiter ist dies jedoch, so weit wir empirisch und theoretisch wissen, in relevanten Fällen wahrscheinlich nahezu ausgeschlossen (siehe auch (5.7)). Das heißt, wir sind immer mit Beschränkungen aus einer der beiden Klassen konfrontiert (5.7.2). Unter diesen Voraussetzungen stellt sich die Frage, unter welchen Bedingungen Menschen überhaupt gerechtfertigt eine Erklärung akzeptieren, wenn sie doch die Bedingungskonstellation $x_1 \dots x_n$ niemals vollständig überblicken können.

5.8.1 Institutionen

Die (soziale) Erkenntnistheorie und Soziologie gibt hier mehrere Antworten auf die Frage, warum Menschen Erklärungen akzeptieren, von denen ich hier zwei wesentliche herausstellen möchte.

Es gibt eine wiederum nicht überschaubare Menge an normierenden Institutionen, die implizit und explizit regulieren, welche Bedingungen $x_1 \dots x_n$ als legitim zur Erklärung von Ereignis Y gelten und welche nicht. Institutionen erfüllen hiermit gewissermaßen die Funktion des Regressstoppers im Spiel des Gebens und Verlangens von Gründen im Kontext unüberschaubarer Komplexität (Gabriel, 2016a; Habermas, 1992, S. 100ff.). In hochkomplexen, modernen Gesellschaften wie der Bundesrepublik gehören hierzu Schulen, Universitäten, Gerichte, Parlamente, Gesetze, Behörden, Gefängnisse und so weiter. Aber auch Freundesgruppen, Vereine, Familien, Beziehungen, sowie soziale und geschlechtliche Identitäten regulieren diese Verhältnisse (Gabriel, 2016a, S. 104). Gesellschaftliche Institutionen und Identitäten erfüllen die Funktion, dass sie Komplexität in modernen Gesellschaften moderieren und reduzieren (Luhmann, 1968, S. 1ff.). Es gibt also Mechanismen, die Erklärungen *als* Erklärungen legitimieren bzw. delegitimieren. Diese wirken hochgradig selektiv, indem sie bestimmte Bedingungen gar nicht erst zulassen, produktiv, indem sie bestimmte Bedingungen erzeugen und entdecken und sanktionierend, indem sie Verstöße bestrafen (Foucault, 1974, 1991).

Historizität von Erklärungen Diese Mechanismen der Regulierung der Legitimität von Erklärungen sind synchron und diachron variabel (Gabriel, 2016a, S. 104). Das heißt, wenn wir den Katalog der Bedingungen unter dem eine Erklärung *als* Erklärung im Rahmen

einer bestimmten Institution gilt, explizieren, dann stellen wir fest, dass diese in Zeit und Raum variieren (Gabriel, 2016a, S. 94ff.), (Gabriel, 2014). Bemerkenswert ist zum Beispiel, wie sich die Erklärung, dass ein Ereignis als Straftat im Kontext des Justizwesens gilt oder ein Ereignis als *Symptom* einer *psychischen Störung* gilt, durch die Zeit wandelt.[46] Diese Historizität wiederholt sich auf der Ebene der Bedingungen und Prädikate. So wandelt sich nicht nur der Kriterienkatalog der Symptome, welche eine psychische Störung als solche identifiziert, sondern auch die Prädikate Symptom und psychische Störung und ihre jeweiligen Verbindungen unterliegen dem historischen Wandel (Foucault, 1973, 1974, 1975).[47]

5.8.2 Anerkennung und Vertrauen

Das juridische Durchsetzen dieser Bedingungen mittels Selektion, Produktion und Sanktion ist noch nicht hinreichend für den epistemischen Zustand des Verstehens als Vertrauen im oben entwickelten rechtsethischen Verständnisses. Bis hierhin ist die Legitimität von Erklärungen lediglich auf das rein (rechts-)positivistische Durchsetzen eines normierenden Apparats reduziert. Auf der Folie des oben skizzierten autonomen Menschenbildes (4.4) benötigen wir als Menschen nicht nur einen Apparat der Durchsetzung der Bedingungen des Erklärens, sondern *gute* Gründe, warum wir diese (und nicht jene oder überhaupt eine Erklärung) akzeptieren sollten.[48] Menschen können in rechenschaftspflichtigen Verhältnissen insofern selbstgesetzgebend (autonom) tätig werden, als dass sie diese Erklärungen *als* Erklärungen für sie geltend anerkennen. Unter Verweis auf das oben skizzierte Modell der Freiheit als *rationaler Persistenz* können wir argumentieren, dass sie diese anerkennen, wenn Menschen diesen Bedingungen vor dem Hintergrund ihrer *praktischen Identitäten*, ihrer Werte, Ziele und Interessen zustimmen können und dass sie sich auf die Gewalten (Justizwesen, Kreditinstitute, Schulen usw.) im Verfolgen dieser ihrer Interessen, verlassen

[46] Wenn Kant nach den Bedingungen der Möglichkeit der Erkenntnis *a priori* fragt, fragt die archäologische und genealogisch-historische Forschung im Ausgang von Foucault nach den historischen und damit auch kontingenten Bedingungen der Möglichkeit der Erkenntnis, kurz nach dem *„historischen Apriori" (Foucault, 1974, S. 17ff.)*.

[47] Am prägnantesten wird dies am treffenderen Titel des französischen Originals von *Die Ordnung der Dinge – Les mots et les choses –* deutlich. In dieser Optik stellt Foucault genau die Frage, wie sich die Verbindung der Worte (*Signifikanten*) und der Dinge (*Signifikate*) im Verlauf der Zeit wandelt.

[48] Der Ausdruck *guter Grund* ist mit notorischen Schwierigkeiten behaftet, die hier nicht thematisiert werden können. Wesentlich liegt die Problematik in der Frage, welche Gründe denn normativ positiv, d.h. ermächtigend wirken bzw. negativ, das heißt entfremdend und paternalistisch zu verstehen sind. Es gibt gute Ansätze diese Problematik zu beantworten, doch hier müssen wir die Existenz dieser Distinktion einfach voraussetzen.

können (Knappik, 2013, S. 102ff., 130ff.).[49] Und an dieser Stelle komplementieren sich die rechtsethische und erkenntnistheoretische Besprechung des Verstehens als Vertrauen. Vor dem Hintergrund der (Hyper-)komplexität des Wirklichen bedeutet Verstehen im Kontext moderner Institutionen, dass die Tatsache der Grenzen der Erklärbarkeit im Sinne der Enthüllung der erklärenden Bedingungen (5.7) anerkannt und toleriert werden. Das heißt die Erkenntnis über die Grenzen der Erklärbarkeit selbst sollte anerkannt werden. Es braucht einen Prozess des Sich-darauf-verlassen-Könnens, dass die Institutionen ihren Auftrag im Interesse der Menschen unter Bedingungen der Ungewissheit gerecht werden. Und der Name für diesen Prozess ist *Vertrauen* (Luhmann, 1968, u. a. S. 34f., 51f.). „Vertrauen braucht man zur Reduktion einer Zukunft von mehr oder weniger unbestimmt bleibender Komplexität." (Luhmann, 1968, S. 19) Und hier werden bestimmte spezifizierbare Bedingungen relevant, die Menschen an vertrauenswürdige Institutionen stellen.[50] Diese Bedingungen lassen sich aus den einschlägigen Texten der Rechtsstaatsgeschichte ablesen, wie zum Beispiel der französischen Unabhängigkeitserklärung, dem Code Civil oder dem Grundgesetz, als auch aus den Erkenntnissen der (Rechts-)Soziologie, Sozialphilosophie, der Politik- und Rechtswissenschaften (Epping, Lenz & Leydecker, 2024; Luhmann, 1969; Weber, 1922, S. 55ff.). Durch diese Anforderungen, die unter der Bedingung genuiner Komplexität an Institutionen gestellt werden, spiegeln Institutionen die Eigenschaften des oben skizzierten Menschenbilds (4). Zu nennen sind hier vor allem die Folgenden:

- Kohärenz: Institutionen sollten sich nicht widersprechen und konsistent mit den eigenen Zielen, Werten und Aufträgen agieren.

- Gleichheit: Sie sollten wesentlich Gleiches gleich behandeln und wesentlich Ungleiches ungleich. Dies leitet sich auch aus dem Anspruch nach Kohärenz ab.

- Kontinuität: Institutionen sollten Kontinuität in ihrem Agieren aufweisen. Das heißt die Maßstäbe, die zum Beispiel Erklärungen als solche etablieren, sollten sich nicht sprunghaft, zufällig oder gar willkürlich ändern[51], es sei denn, diese Modifikationen

[49] Dieser Aspekt der Anerkennung von Institutionen ist deutlich komplexer als hier besprochen werden kann, aber die Anerkennung kann als eine fundamentale Dimension eines wertegeleiteten Rechtsstaats angesehen werden (Honneth, 1992).

[50] Luhmann arbeitet klar heraus, dass der Mensch ohne ein gewisses Grundvertrauen wider der Komplexität der Wirklichkeit nicht lebensfähig wäre. „Alles wäre möglich. Solch eine unvermittelte Konfrontierung mit der äußersten Komplexität der Welt hält kein Mensch aus." (Luhmann, 1968, S. 1f.)

[51] In diesem Sinne wäre Willkür zu definieren, als die ungerechtfertigte, nicht-erklärte Entscheidung bestimmte Maßstäbe zu verschieben.

sind gerechtfertigt.

- Transparenz, Erklärbarkeit und Rechenschaftspflicht: Institutionen sollten ihr Agieren begründen können.

- Korrigierbarkeit: Wenn Institutionen fehlerhaft sind, sollten sie korrigierbar sein.

Diese Prinzipien sind idealiter auf die meisten Institutionen anzuwenden, die *a.)* mit hyperkomplexen Prozessen zu tun haben, die Menschen betreffen und *b.)* die Rechte dieser Menschen zu schützen den Auftrag haben.[52] Wie wir wissen, werden die Institutionen diesem Auftrag oftmals nicht gerecht, aber diese Bedingung stellen wahrscheinlich das höchste Maß an Übereinkunft der Institutionen mit den rechtsethischen Prämissen dar, auf denen die (auch digitale) Gesetzgebung beruht (4). Psychologisch und ethisch ist an dieser Stelle noch zu ergänzen, dass die Kontinuitätsbedingung des Verstehens eine notwendige temporale *und* soziale Dimension hat. Vertrauen ist nicht einseitig auf einen Bedingungskatalog zu reduzieren, sondern entfaltet sich erst als anerkannte Institution in der Zeit, die sich durch das Nachkommen ihrer Rechenschaftspflicht Vertrauen erarbeitet haben (Luhmann, 1968, S. 9). „Wer Vertrauen erweist, nimmt Zukunft vorweg. Er handelt so, als ob er der Zukunft sicher wäre. Man könnte meinen, er überwinde die Zeit, zumindest Zeitdifferenzen.“ (Luhmann, 1968, S. 9) Vertrauen ist eine gelebte Praxis. In anderen Worten: Vertrauen muss einerseits technisch-institutionell hergestellt, andererseits zeitlich durch soziale Praxis erworben werden. Letztlich befinden Individuen und Kollektive darüber, ob sie einer Institution (Ämtern, Verfahren, Parlamenten und so weiter) vertrauen und dieses Ereignis ist ebenfalls fundamental nicht-algorithmisierbar.

5.9 Sicherheit als Verstehen und Vertrauen

Wissen: Aus der bisherigen Analyse können wir auf ein weiteres Differenzierungsmerkmal schließen, welches im Vergleich zum Zustand des Wissens und des Vertrauens deutlich wird. Ohne nun auf die Schwierigkeiten des sogenannten analytischen oder kriteriellen Wissensbegriff einzugehen, stellt dieser Wissen als wahre, gerechtfertigte Überzeugung vor. Wenn ein Subjekt S weiß, dass p (zum Beispiel Grund für die Ablehnung des Kredits),

[52] In gewisser Weise lässt sich die gesamte Geschichte gesellschaftlicher Institutionen, als auch des modernen Rechts- und Verfassungsstaats als ein Versuch lesen (Hyper-)Komplexität und dem damit verbundenen Dissens zu managen (Gabriel, 2020a; Luhmann, 1968, § 13).

dann

1. weil S überzeugt ist, dass p,

2. p ist wahr und

3. S ist gerechtfertigt in der Annahme, dass p (Ayer, 1956, S. 34).

Doch wie üblich wird hier, wie in der ganzen Tradition die Rechtefertigungsbedingung, als entscheidende (wenn auch nicht einzige) Schwachstelle identifiziert (Gabriel, 2016a, S. 71). Wenn S p rechtfertigen kann, muss dieser Zustand noch nicht Autonomie-fördernd im obigen Sinne sein. Knappig gibt hierfür folgendes Beispiel. „[...] [V]iele Menschen haben zum Beispiel eine wahre und gerechtfertigte Meinung dahingehend, dass aus der speziellen Relativitätstheorie die Gleichung $E = mc^2$ folgt, ohne genügend von der speziellen Relativitätstheorie zu verstehen, um zu wissen, warum dem so ist. Gewöhnlich würden wir die fragliche Überzeugung dennoch als gerechtfertigt und als Fall von Wissen gelten lassen.“ (Knappik, 2013, S. 377) Auf unser Thema gewendet könnte S die wahre und gerechtfertigte Überzeugung haben, keinen Kredit aufgrund der statistischen Ableitungsregeln eines KI-Modells erhalten zu haben, ohne wirklich diese Ableitungsregeln (die involvierte Logik) und den Beitrag der einzelnen Features zu *verstehen*. Weiterhin könnte S auch aus den sachlich richtigen, doch moralisch falschen Gründen überzeugt sein, dass p. Wenn sich das Modell beispielsweise diskriminierend verhält bzw. sich die Gewichtung einzelner Features vor dem Hintergrund geltender normativer Maßstäbe nicht rechtfertigen lässt. Wir können auch etwas Wahres wissen, die Gründe sind zwar wahr, aber die falschen (d.h. ethisch und/oder juristisch nicht zu rechtfertigende) Gründe. „Das eigentliche Ziel epistemischer Tätigkeit– das, was wir an deren Leistungen wertvoll finden– besteht demnach nicht darin, dass wir etwas wissen, sondern darin, dass wir etwas verstehen.“ (Knappik, 2013, S. 376)

Verstehen und Vertrauen Verstehen können wir dem Gegenüber in diesem Theorierahmen so definieren, dass das Subjekt in einen epistemischen Zustand ist, welcher die Autonomie und Autorenschaft ausweitet. Das bedeutet S hat *gute* Gründe für die Annahme, dass zum Beispiel eine automatisierte Kreditanalyse zu einer entsprechenden Prognose kam. Unter den oben diskutierten Bedingungen kann dies zum Beispiel bedeuten, welche Variablen hätten um wie viele Einheiten höher bzw. niedriger sein müssen, um eine positive bzw. negative Entscheidung zu erhalten. Oder es kann S dargelegt werden, welche

Variable hinreichend für die gegebene Prognose war. In diesem Zuge sollte S auch die Nicht-Diskriminierung in S konkreten Fall nachvollziehen können, dass zum Beispiel die Variable Geschlecht oder Hautfarbe weder notwendig, noch hinreichend für die Prognose war. Nehmen wir an diese Voraussetzungen seien alle erfüllt. Dann wäre das Ziel, dass die Erklärung für S Teil des Prozesses rationaler Transformation wird, in dessen Ergebnis S sich die Erklärung als Teil des Selbst zu eigen machen kann. Im idealen Ergebnis kann S sich mit dieser Erklärung auch *normativ* identifizieren. Das beutetet der Modelloutput ist unter gegebenen moralischen und rechtlichen Normen *gerechtfertigt*. Selbstverständlich setzt dies voraus, dass die Modelle, Datensätze und die Entscheidungsfindung diesen Anforderungen gerecht wird, was wiederum Aufgabe der Verantwortlichen, vor allem der Betreibenden in diesem Zusammenhang ist. In diesem Sinne bedeutet Verstehen als Vertrauen die Möglichkeit der „Maximierung rationaler Kohärenz“ (Knappik, 2013, S. 341). „Ereignisbeherrschung und Vertrauen sind mithin nicht lediglich funktional äquivalente, einander substituierbare Mechanismen der Reduktion von Komplexität; steigt die erfassbare Komplexität möglicher Ereignisse, dann müssen sie beide komplementär und nebeneinander stärker beansprucht werden.“ (Luhmann, 1968, S. 19) Autonomie entwickelt sich in diesem Bild wesentlich zu einer epistemischen Tätigkeit und daher ist sie notwendig mit dem Problem der Interpretierbarkeit verschränkt.[53]

Sicherheit Und sobald wir in Interaktionen mit Technologien und Institutionen in ein solches Verhältnis eintreten, welches durch den Nexus aus Rechenschaftspflicht, Konsistenz und Persistenz zur Autonomieförderung gekennzeichnet ist, *dann* können wir von Sicherheit sprechen. Sicherheit, Verstehen und Vertrauen sind demnach begrifflich dahingehend aufeinander bezogen, dass wir uns auf die Autonomie-fördernde Einrichtung dieser Relationen *verlassen* können sollten, uns eben sicher sein sollten. Sicherheit bedeutet, dass sich-darauf-verlassen-Könnens der Autonomie-fördernden Wirkung einer Erklärung im Verhältnis zu Institutionen, Verantwortlichen und Maschinen unter Bedingungen der *Hyper-Komplexität*. Normativ können wir demnach den Verlust der Möglichkeit von Verstehen und Vertrauen damit als ein Verletzen der informationellen Selbstbestimmung und Integrität einer Person verstehen.

[53] „Das Denken im engeren Sinne wird dabei insgesamt als Prozess verständlich, in dem die in Anschauung und Vorstellung gegebenen Inhalte transformiert werden, was eine Leistung der epistemischen Befreiung und der Konstitution eines rationalen epistemischen Selbst darstellt.“ (Knappik, 2013, S. 24)

6 Nicht-Interpretierbarkeit als epistemisches Risiko

Im vergangenen Kapitel haben wir einen allgemeinen theoretisch-analytischen Rahmen für das Problem Interpretierbarkeit als Verstehen und Vertrauen entwickelt und erste Probleme anhand dessen aufgezeigt. Unter (5.7) wurde auch bereits die These eingeleitet, dass die Komplexität der Wirklichkeit eine prinzipielle Grenze für ein instrumentelles Verstehen markiert. Wir sind im Stande theoretisch, empirisch und formal eine Reihe von Prozessen zu identifizieren, die komplexer sind, als dass sich die erklärenden Variablen in zusammengenommen notwendige und hinreichende Bedingungen gliedern lassen, sodass diese das zu erklärende Ereignis determinieren. Die Prämisse dieses Teils lautet folglich: Eine deterministische Erklärung, wie unter (5.6) dargestellt, gibt es oftmals nicht. In genau diesem Sinne sind diese Prozesse *hyper*-komplex, das heißt, sie weisen eine Komplexität auf, welche nicht vollständig algorithmisiert werden kann. Nun haben Teile der Komplexitätsforschung in den letzten Jahren ein umfassendes Modell zur Beschreibung von Komplexität vorgelegt, welches konsistent ist mit dem obigen Modell der Interpretierbarkeit. Einige der zentralen Autor:innen aus dem letzten Kapitel haben sich auf dieses Modell bezogen bzw. kooperativ an dessen Entwicklung mitgewirkt.

In diesem Kapitel geht es darum a.) das Modell zu skizzieren und in den theoretischen Rahmen zu integrieren, um dann b.) einen Zweig dieser Forschung auf künstliche neuronale Netze anzuwenden. Daraus lassen sich dann einige grundlegende Aspekte über c.) ihre Komplexität und d.) ihre mangelnde Interpretierbarkeit sowie e.) ihr daraus resultierendes Sicherheitsrisiko ermitteln. Methodisch ist noch vorauszuschicken, dass wir hier nur eine reduzierte Betrachtung einiger allgemeiner Eigenschaften von künstlicher neuronaler Netze anstreben. Dies ist darin begründet, dass es für die Betrachtung der Problematik nicht nötig ist, auf komplexe Architekturen wie Transformer oder Convolutional Neural Networks zurückzugreifen.[54] Bereits anhand der allgemeinsten Eigenschaften von künstlichen neuronalen Netzen lassen sich diese Probleme aufzeigen, die sich dann anhand komplexerer Systeme eskalieren lassen (siehe Diskussion (8.2.1)).

[54] Obgleich dies ein interessantes Projekt wäre, aber den Rahmen dieser Arbeit sprengt (siehe auch Diskussion (8.2.1)).

6.1 Ontologie der Komplexität

Das hier veranschlagte Modell der Komplexität erforscht diese in einem anspruchsvollen, ontologischen Projekt.[55] „The aim is therefore an integrative view to show how this holds in all sciences including chemistry and physics, and is of particular significance in understanding digital computers, life and brain.“ (Ellis, 2016, S. viii) Das heißt es wird versucht eine umfassende Modellierung der (empirischen) Realität vorzulegen. Im Folgenden wird eine maximal reduzierte Skizze des Grundmodells beschrieben. Für eine ausführliche Darstellung siehe das Material im Appendix und insbesondere (Ellis, 2012, 2016; Voosholz & Gabriel, 2021).

Komplexität Das allgemeine Modell beschreibt Komplexität als ein netzwerkartiges Interagieren *strukturierter*, *hierarchischer Module* (Ellis, 2012). Die vielfältigen empirischen Prozesse, wie eine automatisierte Kreditrisikoanalyse, das Verhalten von Gasen, das Ausführen von Computerprogrammen, die Reproduktion von Zellen, die Evaluation von Gütekriterien oder das Schreiben einer Abschlussarbeit werden durch Subprozesse, sogenannte Module konstituiert, die insgesamt energie- und recheneffizienter sind als der Gesamtprozess (Ellis, 2012). Jedes Modul lässt sich wiederum in Submodule aufteilen, bis hypothetische Basismodule erreicht sind, die die elementarsten Operationen ausführen. Aufgrund ihrer erhöhten Interaktion und Energiedichte lassen sich Module zu Ebenen zusammenfassen. Nach ihrer Komplexität lassen sich die Module als aufsteigende Hierarchie darstellen.[56] Der Output der Module einer Ebene ist der Input der Module der höheren Ebene. Jede Ebene ist notwendig durch die sie spezifizierende Gesetzmäßigkeiten und Variablen gekennzeichnet. Unter dieser Voraussetzung können die Eigenschaften höherer Ebenen (ihre Gesetzmäßigkeiten und Variablen) zum Beispiel durch die Grobkörnung (coarse-graining) von Variablen auf niedrigeren Ebenen entstehen (Ellis, 2016, S. 12).

[55] Dabei ist die theoretische Motivation, die Feststellung, dass die Physik sämtlicher Komplexität zugrundeliegt, aber ein reduktionistischer Physikalismus nicht überzeugt, da es infinit viele nicht-physikalische Phänomene gibt, die kausal messbar sind. „Physics underlies all complexity, including our own existence. How is this possible? How can our own lives emerge from interactions of electrons, protons and neutrons?“ (Ellis, 2012)

[56] Es ist zu diskutieren, ob eine hierarchische Betrachtung wirklich notwendig ist oder ob nicht eine sogenannte flache oder eine neutrale Ontologie der Wirklichkeit und ihrer Irreduzibilität noch viel eher gerecht wird. „Existence. The different levels are all real, each existing with causal powers in its own right, because [...] they each have determinable effects on the levels above and below them. No level is more real than any other.“ (Ellis, 2016, S. 90). Zur Verteidigung eines nicht-hierarchischen, sogenannten neutralen Realismus (nicht zu verwechseln mit einer flachen Ontologie) siehe (Gabriel, 2016c, S. 356ff.). Für einen ähnlichen Ansatz, den der biologischen Relativität siehe (Ellis, 2016; D. Noble, 2017).

Die Variablen zum Beschreiben eines Gases entstehen durch das statistische Mitteln von Eigenschaften auf niedrigeren Ebenen. Zum Beispiel entstehen der Druck und die Dichte eines Gases aus der zugrunde liegenden molekularen räumlichen und Geschwindigkeitsverteilung. Dies ist der Prozess der Grobkörnung (coarse-graining) zugrundeliegender Prozesse (Ellis, 2016, S.10f, S. 96). Die Freiheitsgrade der zugrunde liegenden Einheiten verschiedener Ebenen, von Molekülen bis hin zu Quarks, verschwinden gewissermaßen auf dieser Ebene. Das Gegenteil ist die Feinskalarisierung (fine-graining) – wir betrachten die Situation auf feineren und feineren Skalen, um zum Beispiel das Verhalten einzelner Moleküle genauer zu untersuchen (Ellis, 2016, S. 106). In der grobskalierten Ansicht gibt es jedoch nicht genügend Informationen, um die Eigenschaften tieferer Skalen genau aufzulösen. Auf dieser Ebene bleiben zum Beispiel die Geschwindigkeit und Flugrichtung, sowie die Schwingungs- oder Rotationszustände einzelner Moleküle verborgen. Dies ist die notwendige Bedingung dafür, dass es die höherstufigen Variablen überhaupt gibt. Die inter- und intra-hierarchischen Prozesse der einzelnen Ebenen und Module müssen partiell voneinander isoliert und gekapselt sein, zum Zwecke der Informationsverbergung und Energieeffizienz (Ellis, 2016, S. 41). Da jede Ebene durch ihre je eigenen Variablen und Gesetze gekennzeichnet ist, bedarf es für jede Ebene eines eigenen Sprach- und Methodenkorpus, um diese zu untersuchen, worin letztlich die Pluralität der Wissenschaften gründet. Daher können wir die unterschiedlichen Ebenen als Gegenstandsbereiche von Disziplinen betiteln, zum Beispiel Teilchenphysik, Atomphysik, physische Chemie aber auch Sozial-, Geistes-, Humanwissenschaften und so weiter (Ellis, 2016, S. 8ff.).

Verallgemeinern wir dieses Beispiel auf die gesamte wissenschaftlich messbare Realität, können wir uns die daraus resultierende Hierarchie vereinfacht wie in der Abbildung vorstellen.

6.1.1 Bottom-up-Reduktionismus

Eine solche hierarchische Ontologie stellt letztlich die Motivation und ein Modell für den weitverbreiteten Reduktionismus bereit. „Some physicists and philosophers claim that from a fundamental viewpoint, higher levels are ‘nothing more than’ an aggregation of lower level phenomena, i.e. they will all emerge by coarse-graining.“ (Ellis, 2012) Während die Dynamik auf niedrigerer Ebene abläuft, zum Beispiel die Diffusion von Molekülen durch ein Gas, ändern sich die entsprechenden grobskalierten Variablen auf höherer Ebene

Table 1.1 The basic hierarchy of structure and causation for inanimate matter (*left*) and for life (*right*) as characterized by academic disciplines

	Inanimate matter	Living matter
Level 10	Cosmology	Sociology/Economics/Politics
Level 9	Astronomy	Psychology
Level 8	Space science	Physiology
Level 7	Geology, Earth science	Cell biology
Level 6	Materials science	Biochemistry
Level 5	Physical chemistry	Chemistry
Level 4	Atomic physics	Atomic physics
Level 3	Nuclear physics	Nuclear physics
Level 2	Particle physics	Particle physics
Level 1	Fundamental theory	Fundamental theory

Abbildung 1: Hierarchische Meta-Physik

Quelle: Darstellung nach Ellis (2012).

als Folge der Änderung auf niedrigerer Ebene, zum Beispiel wird eine ungleichmäßige Temperatur zu einer gleichmäßigen Temperatur (Ellis, 2016, S. 96ff.). Auf dieser Basis kann dann das Modell eines evolutionären Reduktionismus Fuß fassen, welcher nach dem Prinzip „Von den Bakterien zu Bach“ alle Module höheren Ebenen, von Molekülen, über Organismen bis hin zu Sozialverhalten und politischen Organisationen deterministisch aus dem jeweils tieferen Ebenen deduziert (Dennett, 2017). Die Einheiten tieferer Skalen, biochemische und physikalische Prozesse etwa, würden dann durch unseren Organismus, welcher an seine biologische Nische adaptiert ist, grobkörnig als einfache phänomenale (Farben, Geruch, Geschmack usw.) repräsentiert.[57] „You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules.“ (Crick, 1995) Das phänomenale Bewusstsein, das heißt unser qualitatives Erleben der Wirklichkeit im Medium unserer Sinnesmodalitäten wäre demnach eine Benutzerillusion, die durch die Grobkörnung der eigentlichen physikalischen Prozesse entsteht (Dennett, 2018).[58]

[57] Besonders radikal ausformuliert zum Beispiel in (Frankish, 2016).

[58] Dabei sei betont, dass hier keineswegs behauptet werden soll, dass dies der einzige Weg sei, den reduktiven Naturalismus auszubuchstabieren, ich führe hier nur beispielhaft eine Argumentationslinie vor (Dennett, 1993; Frankish, 2016).

6.1.2 Top-down-Holismus

Der hier veranschlagte Zweig der Komplexitätstheorie legt einiges an Evidenz vor, die den *starken reduktiven Naturalismus* unter Druck setzt, wenn nicht gar falsifiziert (siehe unter anderem (Ellis, 2016; Gabriel, 2016c)). Es gibt vielfältig komplexe Formen der Bottom-up verursachten Grobskalierung von Prozessen, durch die sich die Existenz vieler Phänomene, wie Vogelschwärme, die Muster von Schneeflocken und sogar Schwarmintelligenz erfolgreich erklären lassen (Ellis, 2016, S. 15, 106). Einer vollständigen Reduktion entgegen steht allerdings die Beobachtung, dass es messbare Phänomene (Variablen und Gesetzmäßigkeiten) höherer Ebenen gibt, welche nicht hinreichend auf die Grobkörnung tieferer Variablen zurückzuführen sind. „However, some higher level causally effective variables cannot be obtained in this way, as they are demonstrably not coarse grainings of lower level variables." (Ellis, 2012) Die Eigenschaften tieferer Ebenen erfüllten die *notwendigen Bedingungen* für das Zustandekommen echter Komplexität, aber im Falle echter komplexer Systeme nur sehr selten die *hinreichenden Bedingungen* (Ellis, 2012). Beispiele für solche Phänomene sind die kausale Effektivität von Computerprogrammen, die Homöostase in Organismen oder das menschliche abstrakte Denken und Planen (Ellis, 2016, S. 109). Dabei ist eine wichtige Beobachtung die der Äquivalenzklassen bzw. der multiplen Realisierbarkeit. Das bedeutet, dass der gleiche makroskopische Zustand durch unbestimmt viele mikroskopische Zustände realisiert werden kann. Auf der anderen Seite bedingen die Parameter, die auf höheren Ebenen gesetzt werden, die Informationen, die durch die tieferen Ebenen fließen und verarbeitet werden. „Higher level structural patterns channel causation at lower levels in the system, breaking symmetry and so constraining what happens at those levels. And those constraints, expressed for example in terms of effective potentials characterizing a wiring system or a neural network, lead to many different kinds of interesting behaviour by coordinating behaviour at lower levels." (Ellis, 2016, S. 87) Dies führt zur Beobachtung des Phänomens echter Emergenz. Das heißt Phänomene, deren notwendige Bedingungen durch die Gesetze und Variablen tieferer Ebenen beschrieben werden können, die aber prinzipiell nicht von diesen vorhergesagt werden können, simultan aber messbare kausale Effekte auf jene tieferen Ebenen haben. Dadurch werden die *Symmetrien* zwischen den Ebenen gebrochen. Das heißt auch, dass zum Beschreiben dieser Ebenen ein neues *irreduzibles* Instrumentarium notwendig wird (Ellis, 2016, S. 87f.). Im Falle von Gasen führt die zugrunde liegende atomare Theorie zu den makroskopischen Gasgesetzen, der Thermody-

namik und den thermischen Eigenschaften von Gasen. Die Theorien zur Beschreibung der Ebenen teilen dabei eine Schnittmenge von grundlegenden Erhaltungssätzen, wie denen zur Beschreibung von Masse, Energie und des Impulses, aber auf beiden Ebenen werden neue Beschreibungssätze notwendig, die auf die jeweils andere nicht anwendbar sind (Ellis, 2016, S. 119f.).[59]

Knapp zusammengefasst gibt es zwei Bedingungen für den Nachweis dieser Form von Top-down Verursachung (Ellis, 2016, S. 16f.):

- Wenn Änderungen höherer Variablen zu messbaren Änderungen des Verhaltens niedriger Ebenen führt.
- Wenn sogenannte Äquivalenzklassen nachgewiesen werden können. Das heißt Klassen niedriger Ebenen, die die gleichen makroskopischen Zustände realisieren.

Diese Form der Top-down Verursachung wird in der Komplexitätsforschung weiter binnendifferenziert. Ellis unterscheidet zwischen fünf Formen der Top-down-Kausalität, die im Appendix resümiert sind (9.3.2).

6.1.3 Informationsverarbeitende Systeme

Das paradigmatische Beispiel für die *Ontologie der Komplexität*, das heißt die modulare, hierarchische Wechselwirkung zwischen Bottom-up und Top-down Verursachung ist der digitale Computer. Dies liegt an seiner speziellen Eigenschaft, dass digitale Computer als *abstrakte Maschinen* jeden diskreten, informationsverarbeitenden Prozess simulieren können und in diesem Sinne sind sie universelle Maschinen (Ellis, 2016, S. 35ff.).

Die Teile Unter (5.6) haben wir die Standarddefinition eines Algorithmus vorgestellt als „eine *endliche*, *eindeutig* definierte Folge von Anweisungen oder Rechenschritten zur Lösung eines Problems oder zur Durchführung einer Aufgabe“ (H. Müller & Weichert, 2023, S. 16f.).

Als *abstrakte Maschine* bedarf der digitale Computer drei Teile bzw. Komponenten:

- Ein Informationsspeicher (Memory).

[59] Für eine genaue Bestimmung des Konzepts der Emergenz siehe (9.3.2).

- Eine Ausführungseinheit (CPU).
- Eine Kontrolleinheit (Betriebssystem).

Als abstrakte Maschine umfasst die Turing Maschine funktional bereits diese drei Komponenten. Das Band entspricht dem Informationsspeicher (Memory). Der Schreib-/Lesekopf entspricht der Ausführungseinheit (CPU) und der Zustandsregler erfüllt die Funktion einer Kontrolleinheit (Ellis, 2016, S. 36f.). Im Informationsspeicher sind die Programme abgelegt (Software). Die Programme sind eine Folge von symbolisch kodierten Instruktionen, die von der Ausführungseinheit gelesen und ausgeführt werden (Hardware). Turing hat nun gezeigt, dass durch die Änderung der gespeicherten Programme eine solche abstrakte Maschine jedes Programm emulieren kann, welches von einem Computer ausgeführt werden kann (sogenannte Turing-Vollständigkeit) (Ellis, 2016, S. 36).

Soweit wir noch von konkreten Architekturen abstrahieren, müssen noch zwei funktionale Teile ergänzt werden, damit digitale Maschinen das komplexe Verhalten produzieren können, welches wir etwa bei künstlichen neuronalen Netzen oder genetischen Algorithmen beobachten können (Ellis, 2016, S. 37ff.).

Abstraktion: Es ist unmöglich, dass ein Akteur einen programmierbaren Computer in einem Projekt gewissermaßen als monolithischen Block realisiert, der alle Module und ihre Subprozesse simultan überblickt (Ellis, 2016, S. 37). Die Leistungsfähigkeit des Computers hängt davon ab, dass die vielen Aufgaben zergliedert und von Modulen ausgeführt werden (*Divide and Conquer*), deren interne Mechanismen gekapselt und von den anderen Modulen isoliert sind (z. B. CPU, Speicher, Schaltkreise, Verbindungen) (Ellis, 2016, S. 40). Dadurch wird im Informationsfluss ein Flaschenhals erzeugt, der sicherstellt, dass das Gesamtsystem effizienter arbeiten kann (Ellis, 2016, S. 41). Digitale Computer bestehen aus integrierten Schaltkreisen, welche eine CPU enthalten. Diese wiederum besitzt eine arithmetisch-logische Einheit (ALU), aufgebaut aus Transistoren, Dioden, Widerständen und so weiter, welche wiederum aus atomaren Gittern bestehen, durch die sich Elektronen bewegen. Daneben gibt es Schichten von virtuellen Maschinen, die jeweils höhere Programmiersprachen auf der Basis der niedrigeren implementieren. Jede virtuelle Maschine emergiert aus der darunterliegenden. Das Ergebnis sind funktional strukturierte modular aufgebaute Hierarchien, die sich in Gestalt einer komplexen Software

und Hardware Hierarchie realisieren (3) und (2) (Ellis, 2016, S. 46ff.).

Level 7	Applications programs	Data and operations
Level 6	Problem-oriented language level	Classes, objects
Level 5	Assembly language level	Symbolic names
Level 4	Operating system machine level	Virtual memory, paging
Level 3	Instruction set architecture level	Machine language
Level 2	Microarchitecture level	Microprograms
Level 1	Digital logic level	Gates, registers
Level 0	*Device level*	*Transistors, connectors*

Abbildung 2: Software Hierarchie

Quelle: Darstellung nach (Ellis, 2016, S. 48)

Level 7	Global network
Level 6	Local network
Level 5	Computer
Level 4	Motherboard, memory banks
Level 3	CPU, memory circuits
Level 2	ALU, primary memory, bus
Level 1	Logic circuits, registers
Level 0	Transistors, resistors, capacitors
Level −1	*Atomic physics*
Level −2	*Nuclear physics*
Level −3	*Particle physics*
Level −4	*Fundamental theory*

Abbildung 3: Hardware Hierarchie

Quelle: Darstellung nach (Ellis, 2016, S. 47)

Zufall: Die letzte Komponente muss das Problem lösen, dass die Resultate der Rechenschritte eines Computers nicht nur Implikationen dessen sind, was bereits in den initialen Daten vorliegt. Die Lösung für dieses Problem ist das Hinzufügen eines Zufallselements. Das Zufallselement löst den Computer aus der streng deterministischen Logik des Rezept-Modells eines Algorithmus. Dies erlaubt noch keine intelligente Lösung von Problemen, eröffnet aber den Spielraum, dass die Programme sich adaptiv an Eingabedatenströme anpassen. Dadurch können wichtige, strukturierte Informationen aus unstrukturierten oder zufälligen Datenströmen generiert werden. Dies ist die erste Bedingung das maschinell-adaptive Lernen von künstlichen neuronalen Netzen (Ellis, 2016, S. 37, S. 55).

Das Ganze Die Strukturhierarchien für Software und Hardware erlauben es, die kausalen Mechanismen zu verstehen, die von Modulen tieferer Ebenen auf die Module höherer Ebenen wirken und *vice versa*. Ein anschauliches Beispiel liefert ein vereinfachtes Bild der Informationsübertragung in Anlehnung an das OSI-Schichtenmodell mit physikalischer Repräsentation, Netzwerkübertragung und Anwendungsebene (siehe Abbildung (5) und (4)) (Ellis, 2016, S. 53f.). Bottom-up ist der Prozess wie folgt zu verstehen. Auf der untersten Ebene werden die Zeichen etwa nach ASCII als Bitfolge codiert. Die codierte

Information wird dann als Folge elektrischer Signale über Kabel bzw. kabellos durch ein Netzwerk transportiert. Dieser Code wird im letzten Schritt von virtuellen Maschinen ausgelesen, welche bestimmte Photodioden auf einem Bildschirm zum Leuchten bringen. Diese können dann auf der Ebene der Anwendung von dem Nutzenden als Zeichen mit der Semantik natürlicher Sprachen verstanden werden. Im Ergebnis entstehen aus physikalischen Signalen abstrakte, bedeutungstragende Daten. Top-down verläuft der Prozess komplementär. Auf der Ebene der Anwendung werden logische Probleme, etwa das Entwickeln einer Web-Anwendung in maschinenlesbare Befehle übersetzt. Diese werden schrittweise in einfachere Anweisungen zerlegt, kompiliert und schließlich als elektrische Zustände in Logikgattern, Dioden und der arithmetisch-logischen Einheit (ALU) der Hardware realisiert (Ellis, 2016, S. 50ff.). Die logischen Variablen höherer Ebenen lenken damit kausal den Informationsfluss und welche Operationen auf tieferen Ebenen ausgeführt werden bis hin zur physischen Signalebene. Ontologisch bedeutet dies, dass abstrakte Variablen (die Software) die Kausalität physikalischer Prozesse steuert (Hardware) (Ellis, 2016, S. 51). „Modular Hierarchical Structuring of Both Hardware and Software. This enables structured top-down causation in the hierarchy, in particular allowing software patterns to control hardware. In this way, abstract entities have causal effects in the physical universe.“ (Ellis, 2016, S. 37)

Komplexitätsreduktion durch Abstraktion Das Wechselspiel aus Bottom-up und Top-down Verursachung kann nur dadurch gelingen, dass die internen Prozesse der Ebenen voneinander gekapselt und isoliert sind, sodass jede Ebene nur mit einer Abstraktion der umgebenden Ebenen interagieren muss (Ellis, 2016, S. 41ff.). Im Ergebnis ist die Datenverarbeitung des digitalen Computers als hocheffizienter Mechanismus zur Komplexitätsreduktion zu verstehen.

Das Wesen dieser Kausalmechanismen ist die symbolische Struktur, in der Probleme formuliert und entlang der Software und Hardware Hierarchie verarbeitet werden. „This combination of bottom-up and top-down actions enables complex higher level behaviour to emerge from simpler lower level processes, which are orchestrated from above by entering suitable data at the keyboard.“ (Ellis, 2016, S. 53)

	Source		Destination
Level 7	Application	⇒ ⇒ ⇒ ⇒ ⇒	Application
Level 6	⇓ Presentation		⇑ Presentation
Level 5	⇓ Session		⇑ Session
Level 4	⇓ Transport		⇑ Transport
Level 3	⇓ Network	Routers	⇑ Network
Level 2	⇓ Link	Link layer switch	⇑ Link
Level 1	Physical ⇒ ⇒	⇒ Cable/wireless ⇒	Physical

Abbildung 4: Datenkommunikation

Quelle: Darstellung nach (Ellis, 2016, S. 54)

Level 7	Application	Message, HTTP/ SMTP /FTP
Level 6	Presentation	Data compression/encryption
Level 5	Session	Data delimitation, synchronisation
Level 4	Transport	Segments, TCP
Level 3	Network	Datagrams, IP
Level 2	Link	Frames, Ethernet/WiFi/PPP
Level 1	Physical	Individual bits, protocols

Abbildung 5: Hierarchie Datenkommunikation

Quelle: Darstellung nach (Ellis, 2016, S. 54)

Mensch-Computer-Interaktion Ein Analoger Vorgang wiederholt sich auf der Ebene des menschlichen Geistes bzw. im Kontext der Mensch-Computer Interaktion auf der Ebene des Nutzenden (siehe Abbildung (6)) (Ellis, 2016, S. 81ff., S. 351ff.). Der menschliche Geist nutzt zur Komplexitätsreduktion die Fähigkeit, die Wirklichkeit symbolisch abstrakt zu repräsentieren. „Our unique human ability is to be able to use symbolic systems, that is, systems of referencing to things, events, and actions that enable us to make coherent symbolic models of the physical, ecological, and social world around us that represent it reasonably well.“ (Ellis, 2016, S. 354) Begriffe sind ganz wesentlich (aber nicht ausschließlich) eine Abstraktionsleistung mit dem Ziel der Gleichsetzung von Ungleichem. Beispielsweise abstrahieren wir von den unendlich vielen Eigenschaften, die Gleis 1 und Gleis 9 voneinander

unterscheiden, um den Begriff *Gleis* zu bilden (Johnston, 2009, S. 131f.), (Gabriel, 2016c, S. 40f.). Begriffe dienen der Komplexitätsreduktion durch Abstraktion. Diese Abstraktion wird wiederum in der Mensch-Maschine-Interaktion insoweit relevant, da Probleme logisch beispielsweise in der Sprache von Pseudocode oder objektorientierter Programmierung codiert werden können. Diese Codierung lenkt in späteren Schritten der Entwicklung wiederum entlang der Software Hierarchie den Informationsfluss der niedrigerer Ebenen. Programme sind in diesem Sinne *nicht-physische*, aber kausal wirksame, abstrakte Entitäten (Ellis, 2016, S. 71f.). Die Mensch–Maschine-Interaktion ist dabei der zentrale Ort der kreativen Kausalität: Symbolische Modelle (z. B. mathematische Berechnungen, Pseudocode, Graphen) generieren Bewertungen, die als höherstufige Ursachen konkrete physische Prozesse anstoßen (vom Problem, zum Algorithmus bis zur Elektronenbahn in Logikgattern) (Ellis, 2016, S. 50). In dieser Interaktion wird das menschliche abstrakte Denken und Planen kausal wirksam. Alles in allem ist die Mensch-Computer-Interaktion folglich als ein holistisches Ganzes zu denken, bei dem Bottom-up und Top-down Verursachung reziprok wirksam werden (Ellis, 2016, S. 81). „At a higher level, the existence of computers is an outcome of the human drive for meaning and purpose: it is an expression of the possibility space of meanings, the higher levels wherebyweguidewhatactionstakeplace.“ (Ellis, 2016, S. 81)

User level	Specific purpose	Goal
Logical level	Problem structure	⇓
Programme level	Particular programmes	⇓
Data level	Specific data	⇓
Physics level	Hardware	Electrons

Abbildung 6: Kausalitätsfluss digitaler Computer

Quelle: Darstellung nach (Ellis, 2016, S. 50)

6.1.4 Implikationen: Gründerealismus, Normativität und Interpretierbarkeit

Diese kurze Skizze zeigt, inwiefern dieser Zweig der naturwissenschaftlichen Komplexitätsforschung kompatibel ist und mehr noch, die oben entwickelte Anthropologie und ihre ethischen Implikationen stärkt. Insbesondere hervorzuheben, ist die Anerkennung der Pluralität der Gründe bzw. Bedingungen in Konsistenz mit obiger Interpretation des Satzes

vom zureichenden Grunde (5.2) und die Anerkennung dieser Gründe als kausal messbar in Konsistenz mit dem obigen Kausalitätsmodell (5.5). Eine entscheidende Implikation dieser Forschung sind die Konsequenzen der sogenannten abstrakten oder symbolischen Top-down Verursachung (siehe Appendix (9.3.2). Theorien, Pläne und Argumente sind zunächst abstrakte Gedankensysteme. Nehmen wir als Beispiel Newtons Fallgesetze. Zunächst sind diese *deskriptiv*, indem sie die inferentiellen Beziehungen der physikalischen Realität, in diesem Falle der Bewegung von Körpern beschreiben. Doch zudem haben physikalische Gleichungen, sowie jede abstrakt-symbolische Beschreibung, wie eine physikalische Gleichung im Kontext menschlicher Diskurse auch immer eine *präskriptive*, das heißt normative Dimension. Fallible Gedankensysteme sollen die Wirklichkeit so repräsentieren wie sie ist und bieten damit als Beschreibung auch einen *Grund* der für oder gegen diese spricht. Wenn wir Newtons Theoreme konsequent durchdenken, *sollen* wir akzeptieren, dass die Kraft das Produkt von Masse und Beschleunigung ($F = m * a$) ist. Diese einfache Überlegung soll nur die für unser Thema entscheidende Schlussfolgerung veranschaulichen, dass in einer transdisziplinären Anthropologie, die Recht, Philosophie, die anderen Geistes- und die Naturwissenschaften umfasst, die *Affektion durch Gründe* nicht nur eine juristisch-pragmatische Prämisse ist, sondern ontologisch verteidigt werden kann (Ellis, 2016, S. 195ff.), (Deutscher Ethikrat, 2023, S. 146). Laut den hier zitierten Autor:innen der Komplexitätsforschung und Philosophie sind wir wissenschaftlich gerechtfertigt in der Annahme, dass der logische Raum der Gründe, in dem wir uns bewegen und leben, keine Nutzerillusion (oder subjektiver Schein) ist, der wie auch immer, auf einen logischen Raum der Ursachen superveniert.[60] Damit bekommt der normative Ausgangspunkt einer rationalen Autonomie (4.2) und der diese implizierte Gründerealismus (4.4) ein gestärktes Fundament. Die kausale Kraft abstrakten Denkens und entsprechender Institutionen erklärt auf nicht-reduktionistische Weise die Bedeutung von Interpretierbarkeit für unser Selbst- und Weltverhältnis, sowie unsere gesellschaftliche Organisationsform.[61]

[60] Für einen anderen, wissenschaftsorientierten Ansatz siehe (Meyer, 2024). Dabei verfolgt Meyer die Strategie, über die diskursive Vernunftnatur handelnder Subjekte nachzuweisen, dass diese sich prinzipiell nicht rein als empirische Objekte beschreiben lassen. Dieser Ansatz ist wissenschaftsorientiert und explizit nicht metaphysisch und in diesem Sinne auch dezidiert nicht-platonistisch, da er sich an der realen Praxis der Disziplinen orientiert und nicht an einem philosophisch stilisierten Abbild. Es wäre ein lohnenswertes Projekt die ontologischen Verhältnisse zwischen Ellis Platonismus und dem wissenschaftsorientierten Ansatz Meyers genauer zu besprechen.

[61] Diese Schlussfolgerung wird hier nur unzulänglich ausgearbeitet, da sie nicht im Zentrum dieser Arbeit steht. Für die Details siehe (Ellis, 2016, S. 195ff.) *Symbolism and Effectiveness of Thought.*

6.2 Künstliche neuronale Netze als komplexe Systeme

Aufgrund ihrer Netzwerkarchitektur sind ANNs paradigmatisch für Ellis Komplexitätsmodell, denn „genuine complexity can only emerge from networks of causation involving modular hierarchical structures“ (Ellis, 2012). Alle weitergehenden Analysen basieren als Referenzmodelle immer auf der im folgenden als *Standardarchitektur* genannten Architektur (Ernst, Schmidt & Beneken, 2023, S. 823ff.):

- Fully Connected Feedforward-Netzwerk
- Mindestens eine versteckte Schicht
- Nicht-lineare Aktivierungsfunktionen
- Supervised Learning
- Backpropagation Algorithmus
- Eine der gängigen Loss Functions

Aufgrund ihrer Verbreitung und vor allem ihrer Anschaulichkeit referiere ich als Aktivierungsfunktion für die Standardarchitektur immer auf die ReLU, diese wird auch als Hyperparameter für das Evaluationsmodell eingesetzt.

Die genuine Komplexität von ANNs evolviert aus zwei abstrakten Eigenschaften, die bidirektional durch bestimmte Mechanismen der Bottom-up und Top-down Verursachung realisiert werden (Ellis, 2016, S. 75f.). Diese unterscheiden ANNs von deterministischen Algorithmen:

1. Bottom-up: Randomisierte Initialisierung
2. Top: Down: Adaptive Selektion

6.2.1 Adaptive Selektion

Die (pseudo-)randomisierte Initialisierung der Gewichte wirkt Bottom-up von den Verbindungen der einzelnen Einheiten (der Neuronen) auf das allgemeine Netzwerkverhalten, während die Wahl der Hyperparameter (Loss Function, Optimizer, Anzahl versteckter Schichten und Parameter) das allgemeine Verhalten der einzelnen Gewichtsanpassungen

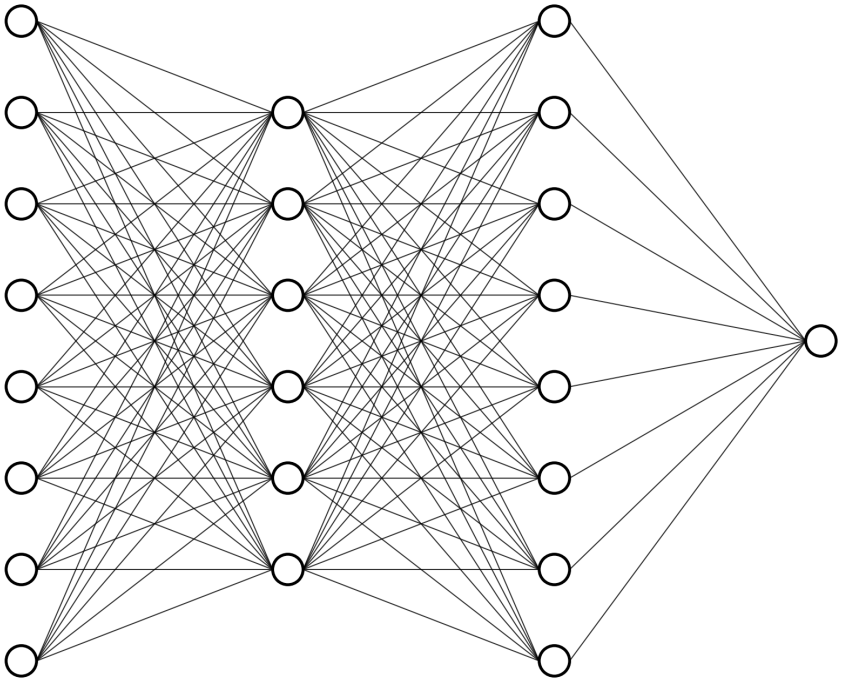

Abbildung 7: Standardarchitektur künstlicher neuronaler Netze

Quelle: Eigene Darstellung

Top-down steuert. Es handelt sich dabei um eine adaptive Selektion, da sich die Gewichtungen, gesteuert durch die Hyperparameter, automatisiert an die eindringenden Datenströme anpassen, indem selektiv Gewichtsmuster verstärkt bzw. geschwächt werden. Im Ergebnis werden Netzwerke von Gewichtungen ausgeprägt, welche sich als Matrizen beschreiben lassen, die aus verrauschten, unstrukturierten Datenströmen strukturierte, latente Variablen aufbauen (Ellis, 2016, S. 57f.).

Formal lässt sich der Prozess der randomisierten, adaptiven Selektion wie folgt darstellen:[62]

$$y_j(t_{i+1}) = j\left(y_1(t_i), \ldots, y_N(t_i), c_j, E\right)$$

Diese Formel modelliert einen zweistufigen Prozess:

[62] Bei der Formalisierung habe ich mich von der formalen Darstellung aus (Ellis, 2016, S. 57ff.) inspirieren lassen.

1. **Variation:** Es wird ein Ensemble von möglichen Systemzuständen erzeugt:

$$\{y_1(t_i), y_2(t_i), \ldots, y_N(t_i)\}$$

2. **Selektion:** Aus diesem Ensemble wird ein bevorzugter Zustand

$$y_j(t_{i+1})$$

ausgewählt, basierend auf einem Selektionskriterium c_j im Kontext der Umwelt E.

Damit beschreibt die Formel, wie aus einer Menge möglicher Zustände mittels eines Selektionskriteriums im Kontext einer Umwelt oder Umgebung einer Zustand aufgebaut wird. Im Ergebnis ist dies der adaptiv-selektive Aufbau neuer Strukturen, der genuin nicht deterministisch arbeitet.

Glossar Formel:

Tabelle 1: Glossar der in der Formalisierung verwendeten Symbole

Symbol	**Bedeutung**
$y_j(t_{i+1})$	Der ausgewählte Systemzustand zum Zeitpunkt t_{i+1}
$y_1(t_i), \ldots, y_N(t_i)$	Das Ensemble möglicher Zustände zum Zeitpunkt t_i
c_j	Das Selektionskriterium, das vorgibt, nach welchen Regeln ein Zustand bevorzugt wird
E	Der Umweltkontext, in dem die Bewertung und Auswahl erfolgt
$j(\cdot)$	Ein Projektionsoperator, der die Selektion durchführt: also aus dem Ensemble den passenden Zustand gemäß c_j und E auswählt

6.2.2 Künstliche neuronale Netze als Approximationsverfahren

Künstliche neuronale Netze können formal als Funktionsapproximatoren beschrieben werden, die eine unbekannt Funktion f sucht, die eine Eingabedomäne X auf eine Zieldomäne Y abbildet: $f : X \to Y$. Und diese Suche ist der Aspekt, der in den meisten ML-Systemen

als Lernen beschrieben wird.[63]

Diese Suchfunktion besteht aus drei Komponenten:

1. Hypothesenraum $\mathcal{H}$: $\mathcal{H}$ umfasst die Menge aller möglichen Funktionen $h : X \to Y$, die das System lernen könnte:

$$\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}$$

   θ sind die Parameter der Hypothese und Θ ist die Menge der zulässigen Parameterwerte.

2. Verlustfunktion $\mathcal{L}$: Die Verlustfunktion misst, wie gut eine Funktion h tatsächlich zu den Daten passt:

$$\mathcal{L}(h, D) = \sum_{i=1}^{n} \ell(h(x_i), y_i)$$

   Hier ist ℓ eine Fehlerfunktion, beispielsweise das quadratische Fehlermaß $\ell(h(x_i), y_i) = (h(x_i) - y_i)^2$ die den Fehler über alle Eingabe-Ausgabepaare summiert.

3. Lernalgorithmus: Ein Optimierungsverfahren, welches die beste Hypothese h^* auswählt, die den Verlust am stärksten minimiert.

$$h^* = \arg\min_{h \in \mathcal{H}} \mathcal{L}(h, D)$$

Die Funktion h^* wird nicht explizit programmiert, sondern aus einer Menge von Trainingsdaten D gelernt:

$$D = \{(x_i, y_i) \mid x_i \in X, y_i \in Y\}_{i=1}^{n}$$

[63] Bei der Formalisierung habe ich mich an gängige Standardwerke der Informatik und KI orientiert wie (Ertel, 2025) und (Russell & Norvig, 2024).

6.2.3 Vereinheitlichung: Adaptive Selektion und Maschinelle Approximation

Das maschinelle Lernen von künstlichen neuronalen Netzen ist ein bestimmter Fall der adaptiven Selektion. Basierend auf einem Selektionskriterium (der Verlustfunktion) wird ein Systemzustand (Funktion) selektiert, welcher das Selektionskriterium befriedigt. Am Ende soll die *beste Lösung* aus einer Menge möglicher Lösungen (Hypothesenraum) gefunden werden. Dabei entspricht der Umweltkontext (die Trainings- und ggf. Nutzungsdaten), das Selektionskriterium (die Verlustfunktion) und der Projektionsoperator (der Lernalgorithmus) der durch den Kontext Top-down wirkenden Verursachung der Makroebene auf die Parameter (Gewichte und Bias) des Netzwerkes, der Mikroebene. Auf der anderen Seite wirken die zufällig initialisierten Werte der einzelnen Parameter, die Verbindungsstärke der einzelnen Einheiten der Mikroebene auf das Gesamtverhalten des Netzwerkes und seiner Interaktion mit der höherstufigen Umgebung, der Makroebene zurück.

Der Grund, warum dies als eine Form von Top-down-Verursachung einzustufen ist, ist, dass die Natur der höherstufigen Umgebung, die Daten und der Umweltkontext entscheidend für die finale Modellstruktur ist. Gemäß dem Kausalitätsmodell (5.5) wäre die resultierende Modellstruktur eine andere, wenn die höherstufige Umgebung modifiziert würde.

Zuordnung Variablen

Adaptive Selektion	**Maschinelles Lernen**
$y_1(t_i), ..., y_N(t_i)$: Zustandsensemble	$\mathcal{H} = \{h_\theta \mid \theta \in \Theta\}$: Hypothesenraum
c_j: Selektionskriterium	$\mathcal{L}(h, D)$: Verlustfunktion
E: Umweltkontext	$D = \{(x_i, y_i)\}_{i=1}^n$: Trainingsdaten
$j(\cdot)$: Projektionsoperator	Lernalgorithmus (zum Beispiel Gradientenverfahren)
$y_j(t_{i+1})$: Ausgewählter Zustand	h^*: Beste Hypothese

6.3 Komplexität und die Nicht-Interpretierbarkeit von künstlichen neuronalen Netzen

Der nun resultierende Mangel an Interpretierbarkeit lässt sich ermitteln, indem wir die soeben beschriebenen Dimensionen Bottom-up und Top-down im Falle der Standardarchitektur abschreiten und ihre Wechselwirkung ausarbeiten.

6.3.1 Hochdimensionalität und In-determinismus

Selbst in einfachen künstlichen neuronalen Netzen werden bereits zehntausende Parameter während des Trainings aktualisiert. Die Parameter des Netzes zu Beginn des Trainings, als auch die Trainingsdaten können als zwei Mengen hochdimensionaler Tensoren repräsentiert werden. Während des Trainings interagieren diese beiden Mengen in Form von Matrizentransformationen. Der Trainingsprozess ist durch zufällige Initialisierung, oftmals zusätzlich Daten-Shuffling und stochastische Optimierung, geprägt. Dies führt auch bei identischen Hyperparametern zu unterschiedlichen Modellzuständen. Auf der anderen Seite ist die Menge der Trainingsdaten zu komplex um ihren Einfluss auf spezifische Parameteränderungen eindeutig rekonstruieren zu können. Das reziproke Zusammenwirken dieser beiden Mengen erzeugt eine Komplexität im Sinne einer nicht vorherzusehenden, In-determiniertheit der Modellstruktur und des Modellverhaltens (Russell & Norvig, 2024, S. 802f.).

6.3.2 Adaption auf Rauschen: Artefakte und Scheinkorrelation

Zu Beginn des Trainings ist die Menge der eindringenden hochdimensionalen Tensoren für das Modell nichts weiter als ungeordnete Datenströme bzw. Rauschen. In diesem Sinne kann man sagen, dass maschinelles Lernen als adaptiv selektiver Prozess Ordnung aus Chaos bzw. Rauschen erzeugt. „The process generates new information that was not there before—or rather, finds information that was *hidden in noise*. That is the general process whereby adaptive selection generates useful information: it finds what is relevant and works from an ensemble of stuff that is mainly irrelevant or does not work, hence allowing a local flow against the general tide of increasing disorder (meine Hervorhebung, J.N.).“ (Ellis, 2016, S. 167) Das ungelernte Modell muss aus diesem Rauschen sinnvolle, das heißt für Menschen versteh- und nutzbare, Muster extrahieren. Die Metriken der Modell-Performance geben zwar Auskunft über die statistische Güte des Modells, aber sie geben keine direkte Auskunft darüber, welche internen Repräsentationen das Modell gelernt hat, um diese Klassifikations- oder Prädikationsperformance zu erreichen. Die internen Repräsentationen müssen keinesfalls notwendig mit den Konzepten korrespondieren, mit denen wir die Welt beschreiben. Es ist möglich und auch empirisch gemessen worden, dass die Modelle in der üblichen Anwendungsumgebung eine gute Performance erreichen, doch die gelernte, interne Repräsentation könnte sich auf für uns vollkommen irrelevante oder

sogenannte spurious Features der Eingabedaten fokussieren. Dabei kann das Modell falsche oder unsinnige Korrelationen als systematische Zusammenhänge ausprägen (Wu et al., 2023; Ye et al., 2024). In diesem Fall wird in der Regel von Artefakten gesprochen. Doch in einem gewissen, unkonventionellen Verständnis erlernen die Modelle immer Artefakte. Das Modell lernt zum Beispiel nicht das semantische Konzept eines „Pferdes“, sondern lernt eine Repräsentation, eine statistische Regularität, die gemessen an den Optimierungsfunktionen hinreichend häufig mit der Zielkategorie korreliert. Im Kreditfall lernt das Modell seine Entscheidungsgrenzen anhand einer hochdimensionalen Repräsentation der für uns verstehbaren Variablen wie Einkommen oder Kreditlaufzeit. Dabei eskaliert das Problem dahingehend, dass das Modell hochdimensionale Konstellationen von Eingabewerten als relevant erachtet. Für den Menschen verständliche Features wie Einkommen oder Kreditlaufzeit verlieren dadurch ihre Rolle als isolierte, kausal interpretierbare Faktoren innerhalb der Entscheidungslogik (Qiu, Kuang & Goel, 2024; Tu et al., 2020).

6.3.3 Komplexität der Umgebung

Neuronale Netze interagieren oftmals mit variablen Umgebungen, das bedeutet neuen Kombinationen von Inputdaten. Aufgrund der Dimensionalität dieser Umgebungen als Ströme von Eingabedaten ist es ab einem gewissen Grad der Komplexität unrealistisch ein Modell auf alle möglichen Kombinationen von Eingabedaten zu testen. Obwohl die Standardarchitektur nach dem Training im Wesentlichen deterministisch arbeitet, das bedeutet, dass bei konstanten Parametern, diese den gleichen Input ($x = x$) zum gleichen Output ($y = y$) transformieren, besteht das Problem darin, dass wir den Datenraum der möglichen Umgebungen nicht vollständig kontrollieren können. Aus diesem Umstand folgt letztlich, dass wir nicht garantieren können, wie sich die Modelle in der Anwendung in neuen Umgebungen verhalten. Dieses kombinatorische Problem lässt sich weiter eskalieren. Mit wachsender Anzahl von Eingabevariablen wächst die Anzahl möglicher Kombinationen exponentiell und damit auch die möglichen Interaktionen aus Parametern und Umgebungsdaten (Perlovsky, 1998). Daher könnte das Modell eine nicht überschaubare Menge an Entscheidungsgrenzen gelernt haben, die wir nicht unbedingt während des Trainings und der Testung alle evaluieren können (Deng et al., 2021; Vankov & Bowers, 2019). Es besteht immer ein Risiko, dass eine Menge an getesteten Umgebungen $\mathbf{u} \in \mathcal{U}$ nicht im sicherheitsrelevanten Sinne repräsentativ ist für $\mathcal{U}$. Dies führt zum Beispiel zu

dem experimentell gut erforschten Befund von sogenannten *adversarial Examples*. Dabei handelt es sich um Wertkombinationen von Eingabedaten, die im Wahrnehmungsurteil eines menschlichen Akteurs nicht zu unterscheiden sind, aber zu unterschiedlichen Outputs führen (Liang et al., 2022). In anderen Worten handelt es sich bei einem vermeintlich identischem x tatsächlich um x', welches wiederum den Output y zu y' modifiziert.

6.4 Sicherheitsrisiko

Oben haben wir eine weite Begriffsdefinition von Sicherheit als Verstehen und (in Interaktion mit Menschen) Vertrauen entwickelt (5.9). Ein entscheidender Aspekt war, dass Sicherheit in Bezug auf Autonomie auch und ganz wesentlich eine epistemische Dimension hat. Ausgehend von diesem erweiterten Sicherheitsbegriff können wir zwei Risikostufen unterscheiden:

1. Ein (nicht notwendigerweise böswilliges) epistemisches Risiko

2. Dieses kann dann zu einem böswilligen Risiko eskalieren

6.4.1 Epistemisches Risiko

Das epistemische Risiko resultiert aus den knapp skizzierten Eigenschaften der Modelle. Wir können die Möglichkeit ihrer vollständigen Sicherheit am Szenario eines maximal intransparenten Modells (welches nicht der Standard sein muss) in Abgleich der Kriterien für vertrauenswürdige Institutionen und Rechenschaftspflicht falsifizieren. Aufgrund der Modellkomplexität können wir nicht garantieren, dass die Modelle wesentlich Gleiches wirklich gleich behandeln und nicht doch diskriminierende, falsche oder unsinnige Entscheidungsgrenzen erlernen, womit sie gegen die Auflage der Kohärenz und Gleichheit verstoßen. Damit zusammen hängt die Beobachtung, dass selbst Modelle, welche in Testung und Evaluation nicht gegen die Auflage der Kohärenz verstoßen, für uns Menschen semantisch unsinnige und kognitiv nicht erfassbare interne Repräsentationen ausprägen können. Im Nexus der Rechenschaftspflicht hat somit eine erklärungsgebende Instanz keine Garantie, dass ein Modell eine Entscheidung wirklich aufgrund der in Testung und Evaluation extrahierten Entscheidungslogik gefällt hat (siehe auch (7.5.13)). Und ihre Korrigierbarkeit hängt letztlich von der Möglichkeit ihrer Transparenz ab. Wenn allerdings die kausale Logik nicht enthüllt werden kann, dann ist auch nicht zu identifizieren, was genau korrigiert

werden soll. Alles in allem kann sich das betroffen Subjekt S_x schlussendlich nicht auf die Autonomie-fördernde Wirkung des Modelloutputs verlassen. Wenn das trialogische Geflecht aus betroffenem Subjekt, verantwortlicher Person und KI-Modell von einem derart intransparenten Modell bestimmt ist, kann die verantwortliche Instanz keine echte Sicherheit im *Geiste des Vertrauens* herstellen (Brandom, 2019). Ein Verstehen im Sinne der *Maximierung rationaler Kohärenz* ist so nicht möglich, da die Modelle selber nicht kohärent im einem anspruchsvollem Sinne sind.

6.4.2 Böswilliges Risiko

All diese Eigenschaften der Modelle können zu einem Sicherheitsverlust führen, ohne dass es eine intentionale, böse Absicht gab. Hier muss niemand jemandem schaden wollen. Doch genau diese Eigenschaften können darüber hinaus von böswilligen Akteuren ausgebeutet werden. Insbesondere können sie die kombinatorische Komplexität dahingehend ausbeuten, dass sie durch gezielte oder auch nur randomisierte adversiale Angriffe (Brute Force) die Modelle zu einem unerwünschtem Verhalten zwingen (BSI, 2024b). Aufgrund der limitierten Möglichkeiten zur Testung und formalen Verifikation der Modelle ist es nicht auszuschließen, dass ein solcher Angriff gelingt (BSI, 2022). Weitergehend können solche Angriffe wiederum aufgrund der Komplexität sehr gut verschleiert werden. Wenn im Falle von Hochrisiko Systemen Modelle in Infrastrukturen integriert sind, in denen sie kritische Bereiche regulieren oder lebensrelevante Güter verwalten, entsteht ein enormes Sicherheitsrisiko (BSI, 2021).

7 Gütekriterien

Der über die letzten Kapitel entwickelte Rahmen bietet einen gewissen Vorzug in Bezug auf die Thematik. Anders als in anderen Arbeiten geht es hier nicht darum, eine spezifische XAI-Methode, wie zum Beispiel attributionsbasierte Methoden (Saliency Maps, LIME), als besonders geeignet für spezifische Anwendungen zu ermitteln.[64] Stattdessen geht es darum, Bedingungen oder Parameter für Methoden zu identifizieren, denen diese gerecht werden müssen, um die hier veranschlagten analytischen, rechtlichen und ethischen Desiderata zu erfüllen (siehe (7.5.10)). Die Gütekriterien wirken im besten Falle wie ein Praxisleitfaden

[64] „It is important for our research community to avoid the one-size-fits-all temptation that there exists a uniquely best way to explain a model.“ (Mothilal & Tan, 2021)

oder zur Inspiration für den Aufbau eines solchen, mit dem Personen aus der genannten Zielgruppe (2.3) Ressourcen bekommen, wie sie den rechtlichen und ethischen Standards unter Zuhilfenahme geeigneter Methoden gerecht werden können. Basierend darauf können Verantwortliche, welche von diesen und ähnlichen Gütekriterien Gebrauch machen, selbst die Methoden auswählen, die am effektivsten mit diesen Kriterien in Deckung zu bringen sind. Zu Anschauungszwecken und als Orientierungshilfe für praktische Anwendung soll hier jedoch beispielhaft eine Methodenkonstellation im Rahmen der Evaluation simuliert werden.

Die Gütekriterien werden nach folgendem Vorgehen entwickelt:

- **Definition:** Als Erstes werden Gütekriterien, wie sie hier verstanden werden, in ihrer Prozesshaftigkeit definiert. Dabei wird eine wichtige Unterscheidung zwischen institutionellen und technischen Gütekriterien vorgenommen (7.1).

- **Institutionelle Gütekriterien:** Dann werden die Parameter für institutionelle Gütekriterien abgesteckt. Hierzu werden die rechtlichen und epistemischen Anforderungen ermittelt, woraufhin die sich damit spezifizierenden Kriterien in die gängigen Sicherheitsstandards nach BSI-Grundschutzkompendium eingebettet werden (7.2).

- **Technische Gütekriterien:** Daraufhin werden die analytischen Desiderata aus dem theoretischen Hauptteil modellhaft auf künstliche neuronale Netze übertragen, um ein Bild einer idealen Erklärung im Anwendungsfall zu gewinnen. Dieses Modell wird dann sogleich problematisiert und zwei mögliche Metriken vorgestellt, die eine partielle Erklärung, das heißt eine Erklärung unter Unsicherheit erlauben (7.3).

- **Die Herstellung von Sicherheit als Verstehen:** Dann wird es Zeit aus den Erkenntnissen bis hierhin die zentralen Desiderata für die Gütekriterien zu synthetisieren (7.4.1), um aus den unterschiedlichen Bausteinen dann einen Ablaufplan der Anwendung der Gütekriterien als Teil des Sicherheitsmanagementsystems zu beschreiben (7.4.2).

- **Evaluation:** Die Schritte dieses Ablaufplans, welche sich auf die technischen Gütekriterien beziehen, werden dann beispielhaft anhand eines recht einfachen Modells zur Kreditanalyse simuliert werden (7.5).

In jedem Schritt sollen die Parameter auf den oben entwickelten Rahmen hin, theoretisch und formal vereinheitlicht und diskutiert werden, wie diese auf die entsprechenden Anforderungen hin modifiziert werden können, um damit eine kohärente Gesamtanalyse anzubieten.

7.1 Definition

Gütekriterien als holistischer Prozess In diesem Kapitel werden ethische, rechtliche, analytische, epistemische und formal-technische Parameter gesetzt, auf die verantwortliche Personen bei der Realisierung von Sicherheit als Verstehen und Verstehen als Vertrauen achten sollten. Die Gütekriterien stellen im Ergebnis *keine Checkliste* vor, die linear abzuschreiten ist. Stattdessen ist das Ergebnis *Desiderata der Anwendung*, sowie ein *Ablaufplan*, wodurch Schritte gewonnen werden, die verantwortungsbewusst als Prozess abgearbeitet werden sollten. Wie herausgearbeitet wurde, gibt es keinen Algorithmus, weder für Verstehen noch für Vertrauen (5.6), sondern nur die ehrliche Arbeit im Trialog mit allen Beteiligten im Geiste des Vertrauens. In diesem Sinne sind die Gütekriterien als holistischer Prozess zu charakterisieren.

Klassen von Gütekriterien Gütekriterien, wie sie hier verstanden werden, sind gewissermaßen Elemente zweier Mengen: technische und institutionelle (bzw. nicht-technische) Kriterien.

1. Technische Gütekriterien (TG): TG sind begriffliche und formal präzisierte, technisch reproduzierbare Kriterien zur Beurteilung sicherer und interpretierbarer KI-Systeme (BSI, 2022), (O. Müller & Lazar, 2024). Gütekriterien, wie sie hier beschrieben werden, vollziehen gewissermaßen eine Übersetzungsleistung: Sie übersetzen die normativen Maßstäbe und den analytischen Rahmen, wie unter (4) und (5.9) herausgearbeitet wurde zum Zwecke ihrer technischen Implementierung und Evaluation. Damit werden normative Ausdrücke wie Sicherheit, Autonomie und Transparenz in einem technischen kontrollierten Umfeld unter gegebenen Beschränkungen (8.2) mess- und reproduzierbar. Konkret können die hier vorgestellten Kriterien als Teil eines Praxisleitfadens dienen, um KI-Systeme auf ihre Sicherheit und Vertrauenswürdigkeit zu prüfen.

2. Institutionelle Gütekriterien (IG): IG könnten auch kurz nicht-technische Güte-

kriterien im weiteren Sinne bezeichnet werden. Dabei handelt es sich um all diejenigen Bedingungen personeller, organisatorischer, infrastruktureller Art, die zusätzlich zu den technischen Gütekriterien realisiert sein müssen. Dabei handelt es sich zum Beispiel um Anforderungen nach EU KI-Verordnung, DSGVO oder dem BDSG. Auch Vorgaben zum Risikomanagement, Audits oder Sicherheitsstandards wie EN ISO oder das BSI-Grundschutzkompendium sind hierzu zu zählen (BSI, 2023; Herd et al., 2024; Raji et al., 2020). In dieser Arbeit wird dieser Ausdruck dahingehend weiter entwickelt, dass er auch die Anforderungen an epistemologischen Kriterien, sowie an Verantwortung und Rechenschaftspflicht umfasst. Diese sind spezifisch so zu verstehen, dass die technischen Verfahren erst durch die institutionellen Bedingungen zur vollen Entfaltung kommen können. Ein scheinbar triviales Beispiel, jedoch mit weitreichenden Implikationen ist die *Empfehlung* des BSI, einfache Modelle mit wenigen Parametern komplexen Modellen vorzuziehen.[65] Beispiele aus dieser Arbeit sind das Einrichten einer Rolle der XAI-Beauftragten oder der Ablaufplan, der beschreibt wie eine Konstellation technischer Gütekriterien integriert wird in das Konstruieren einer Erklärung, die zugleich den epistemologischen Kriterien gerecht wird.

Die Konstellation, die hier entwickelt und beispielhaft evaluiert wird, ist eine Schnittmenge dieser beiden Mengen.

7.2 Institutionelle Gütekriterien

7.2.1 Rechtliche Parameter

Auf dieser Grundlage gilt es nun herauszuarbeiten, welche Parameter für die Entwicklung von Gütekriterien im hier anvisierten Gegenstandsbereich aus der EU KI-Verordnung resultieren. Einige Bestimmungen resultieren direkt aus der KI-Verordnung, andere müssen erst noch erarbeitet werden, um die teils noch unterbestimmten Formulierungen mit Leben zu füllen. Da durch die KI-Verordnung selbst keine exakten Methoden vorgeschrieben sind, muss die Forschung Vorschläge, Standardisierungsverfahren und Praxisleitfäden entwickeln, die nach bestem Wissen und Gewissen im Einklang mit der Verordnung stehen (Herd

[65] „Aus Sicht der IT-Sicherheit sind einfache und transparente Modelle gegenüber großen und komplexen vorzuziehen. Es empfiehlt sich zu prüfen, ob die Parameteranzahl reduziert werden kann oder intrinsisch interpretierbare KI-Modelle (zum Beispiel Entscheidungsbäume), evtl. auch in Kombination, verwendet werden können.“ (BSI, 2021)

et al., 2024).[66]

Eine wichtige Bemerkung ist noch anzufügen. Die KI-Verordnung ist als Ergebnis einer komplexen Verhandlung konfligierender Interessengruppen anzusehen. Vorsichtig formuliert haben nicht alle diese Partikularinteressen, das moralische Fernziel Sicherheit als Verstehen im hier beschriebenen Geiste zu realisieren. Damit soll darauf hingewiesen sein, dass die KI-Verordnung die gesellschaftlichen Güter von denen in (4) die Rede war nicht unbedingt zufriedenstellend schützt. Daher sollten wir die KI-Verordnung als eine Art Leitfaden für den Mindestmaßstab an sichere und vertrauenswürdige KI lesen. In diesem Geiste ist die Erwartung an die folgenden Gütekriterien, dass sie *erstens* mit der KI-Verordnung konsistent sind und *zweitens*, dort wo es angezeigt ist, über sie hinausgehen.

Holistisches Risikomanagement Zunächst sieht die KI-Verordnung ein umfassendes sogenanntes Risikomanagement (im folgenden auch RM) inklusive Konformitätsbewertungsverfahren für den gesamten Lebenszyklus eines Hochrisiko-KI-Systems vor (European Union (EU), 2024, (81), (125)). Dadurch soll die Entwicklung, Inbetriebnahme und Verbreitung in Deckung mit dem Ziel einer sicheren und vertrauenswürdigen KI gebracht werden (European Union (EU), 2024, (123)). Die Bestimmungen zum Konformitätsbewertungsverfahren folgen einem holistischen Ansatz. Das bedeutet, dass das Risikomanagement sich auf den gesamten Lebenszyklus eines KI-Systems bezieht „in einem kontinuierlichen iterativen Prozess [...] der während des gesamten Lebenszyklus eines Hochrisiko-KI-Systems geplant und durchgeführt wird.“ (European Union (EU), 2024, Art. 9(2))

1. Anbieter müssen ihr System einer Konformitätsbewertung unterziehen.

2. Diese Bewertung muss erneut durchgeführt werden, wenn das System selbst oder sein Zweck wesentlich verändert wird.

3. Anbieter von Hochrisiko-KI-Systemen müssen Qualitäts- und Risikomanagementsysteme einführen, um die Einhaltung der neuen Anforderungen sicherzustellen und die Risiken für Nutzende und betroffene Personen zu minimieren, auch nachdem ein Produkt bereits in Verkehr gebracht wurde.

4. Hochrisiko-KI-Systeme, die von Behörden oder im behördlichen Auftrag eingesetzt

[66] Der Autor ist kein Jurist, daher stellt das Folgende keine rechtswissenschaftliche Analyse dar, sondern ermittelt nur einige der wichtigsten Bestimmungen im Bezug auf die Herstellung vertrauenswürdiger KI.

werden, müssen in einer öffentlichen EU-Datenbank registriert werden, sofern sie nicht zu Zwecken der Strafverfolgung und im Bereich der Migration verwendet werden (European Commission, 2024b, 2025; European Union (EU), 2024).

Transparenz Betriebsanleitung, Dokumentation und Aufzeichnungspflichten Aus den Transparenzbestimmungen der KI-Verordnung folgen bereits einige recht klare Anforderungen an die Verantwortlichen. Hier ist ausdrücklich die Aufforderung seitens des Gesetzgebers, Systeme ausschließlich so zu konzipieren, „dass ihr Betrieb hinreichend transparent ist, damit die Betreiber die Ausgaben eines Systems angemessen interpretieren und verwenden können“ (European Union (EU), 2024, Art. 13(1)). Hierzu gehört auch eine umfassende Betriebsanleitung mit Spezifikationen zum Thema Zweckbestimmung, Risiken, Genauigkeitsmetriken und Informationen zur Interpretation der Ausgaben, sowie Darstellung darüber, wie die menschliche Aufsicht realisiert wird (European Union (EU), 2024, Art. 13). Des Weiteren müssen für Hochrisiko Systeme eine technische Dokumentation erfüllt werden, die glaubhaft belegt, wie das Hochrisiko System die Bestimmungen der KI-Verordnung erfüllt (European Union (EU), 2024, Art. 11). Außerdem muss während des Lebenszyklus des Systems ein ständiges Protokoll geführt werden, welches es ermöglicht, sicherheitsrelevante Ereignisse wie Fehlerausgaben aufzuzeichnen und einzusehen (European Union (EU), 2024, Art. 12). Hinzu kommt eine Einschätzung des Maßes an Robustheit basierend auf genutzten Metriken, als auch aller bekannten und vorhersehbaren Umstände, die dieses Maß beeinträchtigen könnten (European Union (EU), 2024, Art. 13, Anhang IV). Die Betriebsanleitung umfasst auch eine Beschreibung der bekannten und vorhersehbaren Folgen der Verwendung, inklusive Fehlanwendungen. Es müssen Kenntnisse über etwaige Risiken bestehen, damit Betreibende hierüber informiert werden können (European Union (EU), 2024, Art. 13).

Folgenabschätzung Das Risikomanagement verlangt eine Analyse bekannter und prognostizierbarer Risiken der Systeme, innerhalb der vorgesehenen Anwendungsfelder (European Union (EU), 2024, Art. 17). Das Risikomanagementsystem versteht sich als ein kontinuierlicher iterativer Prozess, der während des gesamten Lebenszyklus eines Hochrisiko-KI-Systems geplant und durchgeführt wird und eine regelmäßige systematische Überprüfung und Aktualisierung erfordert. Das Risikomanagementsystem soll während des gesamten Prozesses aufrechterhalten bleiben (European Union (EU), 2024, Art. 9). Teil des Risiko-

managementsystems ist das Ergreifen von Maßnahmen zur Minimierung und Kontrolle der in der Folgeabschätzung festgestellten Risiken. „Zur Beseitigung oder Verringerung der Risiken im Zusammenhang mit der Verwendung des Hochrisiko-KI-Systems werden die technischen Kenntnisse, die Erfahrungen und der Bildungsstand, die vom Betreibenden erwartet werden können, sowie der voraussichtliche Kontext, in dem das System eingesetzt werden soll, gebührend berücksichtigt.“ (European Union (EU), 2024, Art. 9(5))

Menschliche Aufsicht Die KI-Verordnung sieht vor, dass Hochrisiko Systeme während ihrer Verwendung von natürlichen Personen beaufsichtigt werden können, mit dem Zweck, effektiv Risiken zu minimieren. Die Anbietende muss die beaufsichtigende Person in die Lage versetzen, das System hinreichend zu verstehen, inklusive Grenzen, Risiken und Anomalien. Um dies zu realisieren, kann die Anbietende entsprechende Mensch-Maschine-Schnittstellen einrichten. Dabei wird auch betont, dass die Maßnahmen zur Herstellung menschlicher Aufsicht dem Autonomiegrad der Systeme gerecht werden müssen. Betont wird auch, dass die Verantwortlichen befähigt werden, „die Ausgabe des Hochrisiko-KI-Systems richtig zu interpretieren [...].“ (European Union (EU), 2024, Art. 14)

Testung, Transparenz und Rechenschaftspflicht Der Teil des Risikomanagement, der die iterative Testung und Evaluation des Models umfasst, sollte Teil des technischen Reports und des fortwährend zu führenden Protokolls werden. Die Testung eines Hochrisikosystems sollte folgende Elemente umfassen:

- Eine umfassende Testung auf *Robustheit:* Dies umfasst Genauigkeit der Ausgaben und Widerstandsfähigkeit gegen interne und externe Störungen (z. B. Fehler, Störungen, Unstimmigkeiten, unerwartete Situationen), sowie Rückkopplungsschleifen bei selbstlernenden Systemen. Hierzu gehören auch der Schutz vor Manipulation der Trainingsdaten (Data Poisoning) und Verhinderung von Angriffen, die zu fehlerhaften Ausgaben führen (Adversarial Examples) (European Union (EU), 2024, (75), Anhang IV).

- *Transparenz* der Metriken: Es müssen exakte Metriken und Wahrscheinlichkeitsschwellenwerte für die Testung vorab bestimmt werden (European Union (EU), 2024, Art. 9, 15).

- *Informations- und Rechenschaftspflicht*: Für unser Thema von besonderer Relevanz ist das in *Artikel 86* beschriebene *Recht auf Erläuterung*. Dies besagt im Wesentlichen, wenn Personen von Ausgaben eines Hochrisiko-KI-Systems betroffen sind, die ihre Grundrechte, Sicherheit oder Gesundheit beeinflussen, hat die Person ein Recht vom Betreiber eine „klare und aussagekräftige Erläuterung zur Rolle des KI-Systems im Entscheidungsprozess und zu den wichtigsten Elementen der getroffenen Entscheidung zu erhalten." (European Union (EU), 2024, Art. 86) Folglich muss zusätzlich zu Metriken der Evaluation und Testung ein Verfahren entwickelt werden, wie Informationen und Erläuterungen aus den genannten Metriken heraus gewonnen werden sollen (European Union (EU), 2024, Art. 86).

7.2.2 Trialog von Verantwortung und Rechenschaftspflicht

Wie oben herausgearbeitet, ist Sicherheit und Vertrauen notwendigerweise auch immer ein dialogischer bzw. diskursiver Prozess. Um all diesen Anforderungen an Transparenz, Informations- und Rechenschaftspflicht gerecht zu werden, bedarf es einer Institutionalisierung dieses Prozesses. In Kontext von Hochrisiko Systemen ist dieser mindestens ein trialogischer Prozess. Der Trialog besteht aus den drei Instanzen:

- dem KI-System KI
- verantwortliche natürliche Personen V (erklärungsgebende Instanz)
- eine natürliche Person, das betroffene Subjekt S_x bzw. die Menge der betroffenen Subjekte S (Erklärung empfangende Instanz)

Das KI-System KI stellt eine Instanz der Komplexität dar, von deren Komplexität S betroffen ist. V muss geeignete Verfahren etablieren, um der Rechenschaftspflicht im Dienste der Sicherheit von S nachzukommen. Hierzu gehört Transparenz über die Gütekriterien, denen dieser Prozess gerecht werden soll. Wenn V beispielsweise den hier entwickelten Ansatz nutzen würde, muss V Verfahren etablieren (Schnittstellen, Saliency, Extraktion von White-Box-Modellen und so weiter), um diesen gerecht zu werden, als auch S darstellen, welche Gütekriterien zur Anwendung kommen und warum. Dabei steht V mittels dieser Methoden in diesem Sinne im Dialog mit KI, um die Komplexität des Modells verstehen zu können. S ist dabei wiederum nicht vollkommen passiv, sondern steht wiederum im

Dialog mit V und gegebenenfalls auch mit KI. Im Ergebnis ist Sicherheit als Verstehen als institutionelles Gütekriterium ein trialogisch-reziproker Prozess.[67]

Dem ist hinzuzufügen, dass das trialogische Modell nur eine vereinfachte Modellierung darstellt, welche von der eigentlichen Komplexität der Situation abstrahiert. In realen Fällen ist die erklärungsgebende Instanz kein homogener Agent, sondern differenziert sich aus in viele Stakeholder, zu denen in einer Organisation die Betreibenden gehören, aber auch Entwickelnde, Aufsichtsgremien, gegebenenfalls auch Behörden. Im Falle einer Klage sind auch Gerichte und Anwält:innen involviert, so wie viele weitere Parteien. Insbesondere sind konfligierende Interessen auf der Seite der erklärungsgebenden Instanz durchaus denkbar. Auch S_x wird hier nur als abstraktes epistemisches Subjekt idealisiert, wobei vollkommen von der leiblichen, geschlechtlichen, sozialen und ethnischen Situation abstrahiert wird. Inwieweit diese Merkmale wiederum die Gütekriterien in Anwendung modifizieren sollten, bleibt einer anderen Arbeit vorbehalten. Auch sollten wir in einer umfangreicheren Arbeit die erklärungsgebende Instanz nicht als sich selbst transparenten Agenten modellieren, bei dem die einzige Intransparenz zwischen dem KI und den V besteht. Stattdessen besteht partielle Intransparenz zwischen den diversen involvierten Akteur:innen. Die erklärungsgebende Instanz ist also durch eine selbstreferenzielle Intransparenz gekennzeichnet. Das trialogische Modell kann jedoch als sinnvoller Ausgangspunkt zur Evaluation und Diskussion genutzt werden, um darauf aufbauend weitere Schichten an Komplexität hinzuzufügen.

7.2.3 KI-Systeme als Institutionen

Weiterhin stellt sich die Frage, welche Eigenschaften von KI-Systemen als Instanz in diesem Trialog wir konkretisieren können. Aus den rechtlichen Parametern aus (7.2.1) sind diesbezüglich drei Elemente mitzunehmen: Robustheit, Transparenz und Informations- und Rechenschaftspflicht. Um diese Bestimmungen in Kongruenz mit dem hier entwickelten Verstehens- und Vertrauensbegriff zu bringen, lässt sich die Institutionenkritik (5.8.1) und die rechtlichen Vorgaben auf KI-Systeme selbst anwenden. Dabei muss diese den genuinen Eigenschaften der Modelle gemäß, abgeschwächt werden. Robuste, transparente und rechenschaftspflichtige KI-Systeme sollten analog zu Institutionen die oben skizzierte

[67] Für eine detaillierte Auseinandersetzung mit den drei Säulen Verantwortung, Rechenschaftspflicht und Transparenz im Rahmen des ART Prinzips (Accountability, Responsibility, Transparency) siehe (Dignum, 2019, S. 53ff).

Anforderungen erfüllen:

1. **Kohärenz und Gleichheit** → Unter hinreichend ähnlichem Input sollte das KI denselben Output liefern.

2. **Kontinuität** → Modellverhalten sollte sich ohne dokumentierte Änderung an Trainingsdaten, Hyperparametern oder Modellversion nicht signifikant verändern.

3. **Rechenschaftspflicht und Transparenz** → Wenn der Output von KI S betrifft, muss dieser auf interpretierbare Input-Features rückführbar und erklärbar sein.

4. **Korrigierbarkeit** → S sollte die Möglichkeit haben, dass V die Entscheidungslogik darlegt und bei Fehlern eine Korrektur vornehmen kann.

Entscheidend ist zu betonen, dass KI-Systeme diese Anforderungen, wenn überhaupt, eben nur als eine Instanz im Geflecht mit den weiteren Gütekriterien erfüllen können. Dabei ist insbesondere das von Menschen betriebene Risikomanagement und die Aufsicht die relevante Dimension. Die Kriterien 1. und 2. müssen von Menschen getestet werden und die Kriterien 3. und 4. von verantwortlichen, fachkompetenten Personen umgesetzt werden. Dies gilt mindestens für die sogenannte enge (bzw. narrow) KI, wie zum Beispiel einen Algorithmus zur Automatisierung von Bewerbungsverfahren. Sobald in einen solchen Prozess leistungsfähigere KI-Systeme integriert sind, wie zum Beispiel die sogenannten *Reasoning-Modelle*, ist unter Umständen die Rechenschaftspflicht und Korrigierbarkeit von dem System bis zu gewissem Grade selbst auszuführen. Doch auch hier ist darauf hinzuweisen, dass die hier veranschlagte Anthropologie eine vollständige Automatisierung des Faktor Mensch im Trialog zur Herstellung von Sicherheit als Verstehen ausschließt.[68] Das heißt das KI-System kann Transparenz inklusive Erklärung und Rechtfertigung über den eigenen Output nicht selbst für die betroffenen Personen herstellen. Sie werden nur in dem Sinne als Institutionen verstanden, dass sie zwar nicht selbst rechenschaftspflichtig sind, wie etwa Betreibende rechenschaftspflichtig sind, aber sie in ein Gefüge der Rechenschaftspflicht eingebunden sind. Datenverarbeitende Systeme werden hier nicht zu moralischen Agenten aufgewertet, sondern diese Bedingungen spezifizieren die Voraussetzungen dafür, dass ein rechenschaftspflichtiger moralischer Agent, zum Beispiel eine Betreibende, seiner

[68] Diese These wird allerdings in dieser Arbeit nur kurz und indirekt verteidigt. Für eine ausführliche Begründung der Irreduzibilität der menschlichen Urteilskraft im Prozess der Rechenschaftspflicht siehe (Gabriel, 2020a, S. 501) und natürlich (Ellis, 2016, S. 310ff.).

Verantwortung nachkommen kann (Coeckelbergh, 2019).

7.2.4 Epistemologische Parameter

Nehmen wir nun einmal an, diese Bedingungen seien im Ansatz erfüllt. Weiter können wir nun die Bestimmungen aus (5.9) nutzen, um den gewünschten epistemischen Zustand von S weiter zu differenzieren.

In dem Geflecht aus Rechenschaftspflicht, Verantwortung und Transparenz haben wir bis hierhin dafür argumentiert, dass eine Erklärung idealiter genau dann die menschliche Autonomie und Autorenschaft eines betroffenen Subjektes S erweitert, wenn

a) S die Bedingungen, die zum Modell-Output geführt haben, verstehen kann.

b) S die Entscheidungslogik des Modell-Outputs transparent gemacht werden kann.

c) S unter der Bedingung kontrafaktischer Kausalität erfahren kann, welche Einflussvariable hätte anders sein müssen, um einen anderen Output zu erlangen.

d) S über die Grenzen der algorithmischen Erklärbarkeit, dem Verfahren der Testung, inklusive vorhersehbarer als auch unvorhersehbarer Risiken informiert wurde und im Bewusstsein dieser Grenzen in diesen Prozess einwilligt.

Letzte Bedingung ist die notwendige, aber eben keinesfalls hinreichende Bedingung für Vertrauen (5.9). Im idealen Falle kann S die Entscheidungslogik *a.)* tatsächlich kausal nachvollziehen, *b.)* ihre Fairness feststellen, *c.)* bei Defiziten (ungerechtfertigte Entscheidungskriterien oder Gewichtung einzelner Variablen) auf Korrektur hinwirken, um *d.)* diese Erklärung nach (4.4) in ihr Selbstverhältnis zu integrieren.

7.2.5 Kontextuelle Sachlogik des externen/internen Modellverhaltens

Wenn wir nur einmal ganz abstrakt diese Kriterien als formalen Rahmen betrachten, dann könnten auch diskriminierende oder vermeintlich unsinnige Erklärungen diesen in Teilen gerecht werden. Nehmen wir noch einmal den Fall der Proxy-Diskriminierung. S_x bekommt dargelegt, dass es aufgrund seines Wohnorts mit der Postleitzahl 70025 und der Haarfarbe Schwarz den Kredit nicht bekommen hat. S_x könnte in diesem Fall partiell erfahren, welche Bedingungen und Entscheidungslogik zum Modelloutput geführt hat. S_x könnte auch

erfahren, dass wenn S_x bei der Bewerbung eine andere Postleitzahl angegeben hätte, den Kredit bekommen hätte. Diese *Erklärung* ist nicht nur diskriminierend, sondern eben auch *sachlogisch* unsinnig. Beides kann jedoch aufgrund der Blackboxnatur nicht prinzipiell ausgeschlossen werden. Damit diese Kriterien für konkrete Institutionen und Einzelfälle wirksam werden können, müssen die Verantwortlichen die Metriken, Schwellenwerte und generellen Anforderungen für Erklärungen und Interpretationen auf den jeweiligen Kontext der Anwendung hin spezifizieren. Ich nenne dieses Kriterium die *kontextuelle Sachlogik*. In Anbetracht der Vielfalt der Anwendungen von Hochrisiko-Systemen werden ganz unterschiedliche Anforderungen relevant (9.1). Zum Beispiel kann das tolerierbare Risiko bei Erklärungen durchaus variieren, wenn wir Systeme vergleichen, die bei der Kreditvergabe assistieren oder Systeme, welche für die Prognose von medizinischen Notfällen genutzt werden (Hong, Haimovich & Taylor, 2018; Mothilal & Tan, 2021). Darauf aufbauend können wir die Erklärungsmodellierung und Interpretation weiter in zwei Schritte differenzieren. Im ersten Schritt wird das *externe Modellverhalten* modelliert. Im zweiten Schritt sollte dann das *epistemische (Rest-)Risiko* mit Bezug auf die *interne Modelllogik* und die *Blackbox Problematik* problematisiert werden. Nehmen wir als Beispiel wieder den Fall Kreditanalyse. Für die Interpretation eines spezifischen Profils können nun folgende Fragen als leitend gelten:

1. *Narrative Plausibilität:* Lässt sich eine Erklärung mit hoher sachlogischer Plausibilität generieren? Das bedeutet beispielsweise:

    - Ist die hohe/niedrige Suffizienz der angezielten Variable sachlogisch überzeugend?

    - Ist es sachlogisch überzeugend, dass die angezielte Variable notwendig/nicht notwendig für den Output ist?

2. *Kontrafaktische Autonomie:* Ist die Erklärung sachlogisch kontrafaktisch ermächtigend? Das bedeutet beispielsweise:

    - Kann der Person transparent gemacht werden, unter welchen kontrafaktischen Szenarien sie einen anderen Modelloutput hätte erreichen können?

    - Sind diese Szenarien sachlogisch fair/unfair? Die Variable *Einkommen* hätte

man unter Umständen noch beeinflussen können, das Geschlecht oder das Alter nicht und letztere sollten auch nicht negativ zum Modelloutput beitragen.

7.2.6 Einbettung nach BSI Grundschutzkompendium

Das BSI-Grundschutzkompendium, sowie die international zertifizierbaren Sicherheitsstandards stellen drei Ziele der Informationssicherheit vor, welche zugleich die Grundwerte der Informationssicherheit sind:[69]

- Verfügbarkeit
- Vertraulichkeit
- Integrität

Die Schäden für diese drei Grundwerte differenzieren sich in den meisten aller Fälle nach folgenden Szenarien (BSI, 2008, S. 105):

- Verstöße gegen Gesetze, Vorschriften oder Verträge
- Beeinträchtigungen des informationellen Selbstbestimmungsrechts
- Beeinträchtigungen der persönlichen Unversehrtheit
- Beeinträchtigungen der Aufgabenerfüllung
- negative Innen- oder Außenwirkung
- finanzielle Auswirkungen

Im Falle von Hochrisiko-KI-Systemen sind oftmals eine Kombination mehrerer oder gar aller genannten Szenarien möglich. Im Zentrum dieser Arbeit liegt dabei insbesondere die Beeinträchtigung des informationellen Selbstbestimmungsrechts wie oben entwickelt (4). Alle drei Schutzziele erhalten durch die hier vorgestellten Analyse Kontur und Tiefe in Bezug auf den Gegenstand KI. Bringen wir einmal kompakt die drei Schutzziele und den hier entwickelten Begriff der informationellen Selbstbestimmung im Bezug auf Interpretierbarkeit von KI-Systemen zusammen: Informationelle Selbstbestimmung bedeutet, dass das

[69] Für Details siehe insbesondere das Grundschutzkompendium und den BSI-Standard 200-2 (BSI, 2008, S. 14).

betroffene Subjekt S_x darauf vertrauen kann, dass die Informationen, die über S_x verarbeitet werden und die daraus resultierenden Modellprognosen folgende Eigenschaften erfüllen. Informationen und Prognosen, aber auch Erklärungen die S_x betreffen, müssen wahr und korrekt sein. Des Weiteren muss S_x darauf vertrauen können, dass diese vertraulich in Bezug auf ethische und rechtliche Desiderata behandelt werden. Hierzu gehören Aspekte wie Zugriffsrechte, Speicherlimitierung, Verschlüsselung und vieles mehr.[70] Aber hierzu zählt auch, dass Vertrauen in eine Informationsverarbeitung, die letztlich die Autonomie von S_x ausweitet und nicht einschränkt. Und damit folgt unmittelbar auch, dass S_x darauf vertrauen kann, dass die Verantwortlichen ihrer Verantwortung in Bezug auf Rechenschaftspflicht, Transparenz und menschliche Aufsicht nachkommen. Und letztlich, muss S_x darauf vertrauen können, dass diese Informationen, Prognosen und Erklärungen nur in einem Maße verwendet werden, die S_x im Vorfeld bestimmt hat und dass ausschließlich S_x und autorisierte Dritte über diese verfügen können, die Informationen gut geschützt sind und S_x die Herausgabe oder Löschung sämtlicher Informationen fordern darf.

Diese Analyse kann nun zur Hand genommen, etwa für die Schutzbedarfsfeststellung. Dabei handelt es sich um einen der notwendigen Schritte zum Aufbau eines Informationssicherheitsmanagementsystems (kurz ISMS). Bei dieser ist immer zu fragen, welcher Schaden für einen dieser Grundwerte entstehen könnte. Diese drei Schutzziele werden nach gängiger Praxis in diesem Schritt vertikal noch einmal nach Schadensauswirkungen in die drei Kategorien *normal*, *hoch* und *sehr hoch* eingeteilt (BSI, 2008, S. 72ff.).

- *normal:* Die Schadensauswirkungen sind begrenzt und überschaubar.
- *hoch:* Die Schadensauswirkungen können beträchtlich sein.
- *sehr hoch:* Die Schadensauswirkungen können ein existentiell bedrohliches, katastrophales Ausmaß erreichen.

Für Hochrisiko-KI-Systeme ist die Schadauswirkungen *ex hypothesi* tendenziell als *hoch* oder *sehr hoch* einzustufen. Dies liegt im Wesen dieser Systeme begründet, deren Prognosen mitunter die Zuweisung von grundrechtsrelevante Güter betreffen, wie der Zugang zu wirtschaftlichen, medizinischen oder sonstigen Ressourcen mit entsprechenden Implikationen

[70] Für Details siehe (BSI, 2008) und insbesondere auch das Bundesdatenschutzgesetz (Bundesministerium der Justiz und für Verbraucherschutz, 2018).

für die informationelle Selbstbestimmung.

7.3 Technische Gütekriterien

7.3.1 Transfer künstliche neuronale Netze

Die vorangegangenen Überlegungen, insbesondere aus (5.9) können wir nun einmal modellhaft formal übersetzen. Wir erweitern im Folgenden die Formalisierung, die von (Mothilal & Tan, 2021) und (Zhang et al., 2021) vorgeschlagen wurde mit ein paar Anleihen aus der formalen Logik und integrieren sie in den hier entwickelten theoretischen Rahmen. Wir betrachten ein abstraktes neuronales Netz $f : \mathbb{R}^d \to \mathbb{R}$, mit dem Input $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ als Eingabevektor und $y = f(\mathbf{x})$ als Modelloutput. Die Variable U beschreibt die Menge aller realistischen Wertkombinationen aus Inputdaten. Für eine konkrete Eingabeinstanz $\mathbf{x}_0 \in \mathbb{R}^d$ ergibt sich der Modelloutput $y_0 = f(\mathbf{x}_0)$. Für einen spezifischen Fall $f(\mathbf{x}_0)$, besteht die Aufgabe darin, diesen Output unter den vorangestellten Desiderata zu erklären.

Minimalkausalität Eine Teilmenge von Merkmalswerten $\mathbf{x}_j = a$ ist minimal kausal für $f(\mathbf{x}_{-j} = b, \mathbf{x}_j = a) = y^*$, wenn die folgenden Bedingungen erfüllt sind:

1. **Existenz:**

$$\exists \mathbf{u} \in \mathcal{U},\ \mathbf{x}_{-j} = \mathbf{b} \text{ mit } f(\mathbf{x}_{-j} = \mathbf{b}, \mathbf{x}_j = \mathbf{a}) = y^*$$

2. **Modalität:**

$$\exists W \subseteq \mathbf{x}_{-j}, \exists \mathbf{w}', \exists a' \neq a \text{ mit:}$$

$$f(W \leftarrow \mathbf{w}', \mathbf{x}_j \leftarrow a) = y^* \quad \text{und} \quad f(W \leftarrow \mathbf{w}', \mathbf{x}_j \leftarrow a') \neq y^*$$

3. **Minimalität:**

$$\nexists\ \mathbf{x}_s \subset \mathbf{x}_j,\ \mathbf{a}_s \subset \mathbf{a},\ \text{während die Bedingungen (1) und (2) für } \mathbf{x}_s = \mathbf{a}_s \text{ erfüllt sind.}$$

In Worten:

1. **Existenz:** Es existiert ein Kontext, in dem die Teilmenge der Merkmalswerte $\mathbf{x}_j = a$ zum Modelloutput y^* führt.

2. **Modalität:** Es existiert mindestens eine Teilmenge von Merkmalen $W \subseteq \mathbf{x}_{-j}$, so dass wenn diese Teilmenge modifiziert wird (W auf w') und sich die Teilmenge der Merkmalswerte ($\mathbf{x}_j \leftarrow a'$ und $a' \neq a$) ändert, sich der der Modelloutput verändert ($y \neq y^*$).

3. **Minimalität:** Es existiert keine echte Teilmenge von x_j, sodass diese Teilmenge die ersten beiden Bedingungen erfüllt, womit es eine kleinere Teilmenge gäbe, die bereits der Minimalkausalität gerecht würde, womit $\mathbf{x}_j$, grob gesagt, bereits zu groß würde.

Notwendigkeit Die Minimalkausalität können wir nun verstärken zur Bedingung der Notwendigkeit. Die Teilmenge $\mathbf{x}_j = a$ ist eine notwendige Bedingung für die Modellausgabe $f(\mathbf{x}_{-j} = b, \mathbf{x}_j = a) = y^*$, wenn (i) sie alle Bedingungen für Minimalkausalität erfüllt und (ii) die Bedingung 2 für Minimalkausaität erfüllt ist, während für W die leere Menge gilt ($W = \varnothing$). In anderen Worten, die isolierte Änderung von $\mathbf{x}_j$ führt zu einem anderen Output: $f(\mathbf{x}_{-j} = b, \mathbf{x}_j = a') \neq y^*$.

Suffizienz Diese Bedingung führt noch nicht zur Hinreichendheit bzw. Suffizienz der Teilmenge $\mathbf{x}_j = a$. Für die Suffizienz benötigen wir die zusätzliche Bedingung:

$$\mathbf{x}_j \leftarrow a \Rightarrow y = y^* \quad \forall \mathbf{u} \in \mathcal{U}$$

In anderen Worten, für $\mathbf{x}_j = a$ muss die obige Aussage in *allen möglichen* Kontexten gelten, damit $\mathbf{x}_j = a$ eine hinreichende Ursache ist.

Ideale Erklärung Eine Teilmenge von Merkmalswerten $\mathbf{x}_j = a$ ist eine Erklärung für die Modellausgabe y^*, wenn:

1. **Existenz:** Es existiert ein Kontext $\mathbf{u} \in \mathcal{U}$, sodass $\mathbf{x}_j = a$ und $f(\mathbf{x}_{-j} = b, \mathbf{x}_j = a) = y^*$.

2. **Notwendigkeit:** Für jeden Kontext $\mathbf{u} \in \mathcal{U}$, in dem $\mathbf{x}_j = a$ und $f(\mathbf{x}_{-j} = b, \mathbf{x}_j = a) = y^*$ gilt: $f(\mathbf{x}_{-j} = b, \mathbf{x}_j = a') \neq y^*$.

3. **Suffizienz:** Für alle Kontexte $\mathbf{u}' \in \mathcal{U}$ gilt: $\mathbf{x}_j \leftarrow a \Rightarrow y = y^*$.

4. **Minimalität:** $\mathbf{x}_j$ ist minimal, d. h. es existiert keine echte Teilmenge $\mathbf{x}_s \subset \mathbf{x}_j$, sodass $\mathbf{x}_s = a_s$ die obigen Bedingungen 1–3 erfüllt, wobei $a_s \subset a$ (Mothilal & Tan, 2021).

7.3.2 Illustration am Beispiel des Perzeptrons

Veranschaulichen wir diese Bedingungen für eine idealisierte Erklärung einmal am vereinfachten Beispiel des Perzeptrons (Ertel, 2025).

Die allgemeine Darstellung des Perzeptrons als Funktion $f(x)$ mit:

$$f(x) = \begin{cases} 1 & \text{wenn } \sum_i w_{ij} x_i + b > 0 \\ 0 & \text{sonst} \end{cases}$$

Instanziieren wir für unser Beispiel das Perzeptron mit drei Eingabewerten und den folgenden Werten für die Parameter und setzen den Bias b auf 0:

$$f(x_1, x_2, x_3) = \begin{cases} 1 & \text{wenn } \sum_i 0{,}4x_1 + 0{,}1x_2 + 0{,}1x_3 \geq 0{,}5 \\ 0 & \text{sonst} \end{cases}$$

Mit dem Eingabevektor:

$$\mathbf{x}_0 = (1, 1, 1) \Rightarrow f(\mathbf{x}_0) = 0{,}6 \Rightarrow y = 1$$

Nun simulieren wir die vier Bedingungen (1. Existenz, 2. Notwendigkeit, 3. Suffizienz und 4. Minimalität) anhand der Fälle $x_1 = 1$ und $x_2 = 1$:

1. Existenz Frage: Existiert mindestens ein Kontext, in dem die Veränderung von x_i zu einem veränderten Modelloutput ($y \neq y^*$) führt?

Fall $x_1 = 1$:

- $x_1 \leftarrow 0: \ f(0, 1, 1) = 0{,}2 < 0{,}5 \Rightarrow y = 0$

Die Änderung beeinflusst das Ergebnis: $x_1 = 1$ ist minimalkausal.

Fall $x_2 = 1$:

- $x_1 \leftarrow 0: \ f(0, 1, 1) = 0{,}2 < 0{,}5$

- $x_2 \leftarrow 0: \ f(1,0,1) = 0{,}5 \Rightarrow y = 1$

In Kombination mit $x_1 = 0$ kann $x_2 = 1$ entscheidend sein. Das heißt, in einem Kontext ist $x_2 = 1$ minimalkausal relevant, aber nicht notwendig.

2. Notwendigkeit Frage: Reicht die Änderung von x_1 aus, um die Vorhersage zu ändern?

- $W = \varnothing$: $\quad x_1 = 1 \Rightarrow y = 1, \quad x_1 = 0 \Rightarrow y = 0$
- $x_1 = 1$ ist notwendig, $x_2 = 1$ nicht $(f(1,0,1) = 0{,}5 \Rightarrow y = 1)$

3. Suffizienz Frage: Ist $x_1 = 1$ allein ausreichend?

- Test: $f(1,0,0) = 0{,}4 < 0{,}5 \Rightarrow y = 0$
- Also: $x_1 = 1$ ist nicht hinreichend
- Kombination $x_1 = 1, x_2 = 1$ ist hinreichend: $f(1,1,0) = 0{,}5 \Rightarrow y = 1$

1. Minimalität Frage: Gibt es eine kleinere Kombination von Werten für x_1, x_2 die bereits die Bedingungen (1-3) erfüllt?

- Für $\{x_1 = 1, x_2 = 1\}$ gilt:
  - In Kombination sind beide hinreichend $(f(1,1,0) = 0{,}5 \Rightarrow y = 1)$.
  - Verkleinert man die Werte oder entfernt einen der Werte
    - $x_1 = 1$: $f(1,0,0) = 0{,}4 < 0{,}5 \Rightarrow y = 0$
    - $x_2 = 1$: $f(0,1,0) = 0{,}1 < 0{,}5 \Rightarrow y = 0$

    dann wird die Ausgabe nicht mehr $y^* = 1$.
- Damit existiert **keine kleinere Teilmenge**, die die Bedingungen (1–3) erfüllt.

7.3.3 Illustration am Beispiel Kreditvergabe

Das Modell der idealen Erklärung können wir noch genauer anhand der Kreditvergabe simulieren. Nehmen wir hierfür wieder das Beispiel Einkommen.

Das Perzeptron $f(\mathbf{x})$ generiert für das Profil von S mit den Werten $\mathbf{x}_0 = (\text{Einkommen} = 2{,}500,\ \text{Alter} = 40,\ \text{Laufzeit} = 24, \ldots)$ die Entscheidung $y^* =$ "guter Kredit".

1. Existenz Es existiert mindestens ein Kontext $\mathbf{u} \in \mathcal{U}$, in dem Einkommen = 2,500 zu y^* = guter Kredit führt.

2. Notwendigkeit Frage: Führt eine isolierte Änderung von x_j zu einem veränderten Output?

- Einkommen = 1,500 $\Rightarrow f(\mathbf{x})$ = schlechter Kredit
- Einkommen = 2,000 $\Rightarrow f(\mathbf{x})$ = schlechter Kredit
- Einkommen = 3,000 $\Rightarrow f(\mathbf{x})$ = guter Kredit

Damit ist Einkommen = 2,500 *notwendig*, da eine Änderung den Modelloutput kippen kann.

3. Suffizienz Frage: Reicht Einkommen = 2,500 alleine hin, unabhängig von den umgebenden Werten, das heißt dem Kontext?

- Einkommen = 2,500, Laufzeit = 24, Sparkonto = mittel $\Rightarrow$ guter Kredit
- Einkommen = 2,500, Laufzeit = 72, Sparkonto = leer $\Rightarrow$ schlechter Kredit

Einkommen = 2,500 ist folglich *nicht hinreichend*, da die Entscheidung von weiteren Variablen abhängt.

4. Minimalität Die Kombination Einkommen = 2,500 und Laufzeit = 24 könnte bereits hinreichend sein, da der Modelloutput in vielen Kontexten stabil bleibt. In dem Fall erfüllt diese Teilmenge die Bedingung der *Minimalität*.

Schlussfolgerung In diesem Beispiel erfüllt Einkommen = 2,500 die Bedingungen von *Existenz* und *Notwendigkeit*, jedoch nicht die *Hinreichendheit*. Eine *minimale Erklärung* könnte durch die Kombination Einkommen und Laufzeit gebildet werden.

7.3.4 Problematisierung und partielle Erklärung

Wie bereits dargestellt, ist eine solche idealisierte Erklärung für künstliche neuronale Netze ab einer bestimmten Komplexität *de facto* unrealistisch (6.2). Dies liegt zentral, wie oben dargestellt, in der kombinatorischen Komplexität begründet. Eskalieren wir die idealisierte Erklärung nun am Beispiel des Evaluationsmodells und des Evaluationsdatensatzes.

Komplexität des Modells Solange wir nur mit einem ein- oder mehrschichtigen Perzeptron interagieren, erlernt das Modell immer eine lineare Hyperebene im Merkmalsraum. Das bedeutet in der Standardarchitektur *ohne Aktivierungsfunktion* können wir unabhängig von der Tiefe des Netzes den Beitrag eines jeden Features zum Modelloutput, das heißt seinen Koeffizienten berechnen (Ertel, 2026, S. 211, 287ff.), (Goodfellow, Bengio & Courville, 2016, S. 155ff.). Daraus folgen zwei bestimmende Eigenschaften für Interpretierbarkeit:

1. *Globalität:* Der *globale* Koeffizient jeder Variable, zum Beispiel Einkommen oder Kredithöhe kann genau bestimmt werden.

2. *Unabhängigkeit:* Der Beitrag jeder Variable ist *unabhängig* voneinander.

Die entscheidende Beobachtung dabei ist, dass wir in der linearen Welt des Perzeptrons *unabhängig* von einem *konkreten* Profil den *globalen* Beitrag einer bestimmten Variable zum Modelloutput kalkulieren können (Saleem et al., 2022; Zhang et al., 2021).

Dieses Bild ändert sich, wenn wir die Modellkomplexität erhöhen (Goodfellow, Bengio & Courville, 2016, S. 166ff.). Im Falle der Standardarchitektur bedeutet dies, dass wir vor allem zwei Hyperparameter verändern:

1. Das Hinzufügen von nicht-linearen Aktivierungsfunktionen

2. Die Anzahl der Parameter und die Tiefe des Modells

Die lineare Merkmalskombination der Eingabemerkmale z_n wird durch Anwendung von ReLU in einen nicht-linearen Output h_n transformiert:

$$\mathrm{ReLU}(z) = \begin{cases} z, & \text{wenn } z > 0, \\ 0, & \text{sonst.} \end{cases}$$

$$h = \text{ReLU}(Wx + b)$$

Durch die Anwendung dieser und anderer Aktivierungsfunktionen (z. B. TanH oder Sigmoid) auf stetig tiefer wachsende Netze verlieren wir genau diese Eigenschaften der Interpretierbarkeit (Goodfellow, Bengio & Courville, 2016, S. 175, S. 193ff.),(Zhang et al., 2021). Dies können wir exemplarisch veranschaulichen, indem wir der Variable *Vermögen* durch das Evaluationsnetz folgen.

Die Variable *Vermögen* trägt jeweils nur den *wertvollsten* Vermögensbestand des Kreditbewerbers, mit den vier kategorialen Ausprägungen:

- A121: Immobilie
- A122: Bausparvertrag/Lebensversicherung
- A123: Auto oder anderes, nicht in Attribut 6
- A124: kein Eigentum

Nach der One-Hot-Codierung werden diese vier kategorialen Ausprägungen der Variable *Vermögen* als binärcodiertes Array mit genau einer positiven Ausprägung (= 1) repräsentiert. Das heißt, die Variable *Vermögen* wird hier auf die einfachste denkbare Weise für das Training abgebildet.

Im *German Credit Data*-Datensatz existieren 13 kategoriale Variablen wie *Status_des_Girokontos* oder *Kreditgeschichte* mit insgesamt 55 Ausprägungen. Durch die One-Hot-Codierung entstehen 55 Dummyspalten. Hinzu kommen 7 numerische Variablen, sodass der Eingabevektor 62 Werte pro Profil umfasst. Daraus ergibt sich eine Eingabematrix der Dimension 1000×62 (Niehus, 2025).

Nennen wir die Transformation, die eine einzige Variable vom Input zum Output durchläuft, einen *Pfad*. Die Anzahl der möglichen Pfade wächst mit der Anzahl der Parameter (Goodfellow, Bengio & Courville, 2016, S. 219ff.). Das kleine Evaluationsmodell mit der Architektur $6-8-1$, also nur 15 Neuronen, besitzt rechnerisch:

$$\underbrace{(62 \times 6)}_{\text{Input–Hidden1}} + \underbrace{(6 \times 8)}_{\text{Hidden1–Hidden2}} + \underbrace{(8 \times 1)}_{\text{Hidden2–Output}} + 15 \text{ Biases } \Rightarrow \text{ 443 Parameter.}$$

Die Architektur umfasst lediglich 14 Neuronen die durch ReLU aktiviert werden können, d. h. sie feuern oder sie feuern nicht. Hinzu kommt noch das Output-Neuron.

Training Während des Trainings werden die Parameter fortlaufend aktualisiert. Für die 14 ReLU-Neuronen existieren 2^{14} mögliche Aktivierungsmuster. Der Zustandsraum (state space, hier symbolisiert durch $\mathcal{S}$ (Russell & Norvig, 2024, S. 67f.)) kann somit beschrieben werden als:

$$\mathcal{S}_{\text{Train}} = \mathbb{R}^{443} \times \{0, 1\}^{14}$$

Der Zustandsraum der während des Trainings durchschritten werden kann ist folglich unendlich groß.

Nach dem Training Nach Abschluss des Trainings sind die Parameter fixiert, sodass lediglich die *diskrete* Kombinatorik der Aktivierungen verbleibt. Das neuronale Netz stellt dann eine deterministische Funktion für jedes Eingabeprofil $\mathbf{x}$ dar. Da das Modell 14 ReLU-Neuronen enthält, existieren weiterhin

$$2^{14} = 16{,}384$$

mögliche Aktivierungsmuster. Jedes Profil aktiviert auf seinem Weg durch das Netz eines dieser Muster.

Aufgrund der *zufälligen* Initialisierung der *kontinuierlichen* Parameterwerte kann sich bei identischen Hyperparametern ein unterschiedliches Modellverhalten ausprägen (siehe auch (Goodfellow, Bengio & Courville, 2016, S. 198ff.)). Entsprechend durchläuft die Variable *Vermögen* bei zwei Modellen, deren Hyperparameter zu Beginn des Trainings identisch sind, niemals die gleiche Transformation. Für ein einzelnes Kreditprofil $\mathbf{x}_0$ ist genau eines dieser Muster aktiv. Da ReLU nur Werte $z_n > 0$ weitergibt, hängt die Weitergabe der Variable

Vermögen von den Werten der anderen Variablen, das heißt vom gesamten Profil, ab. Durch das Setzen negativer Werte auf 0 kann es dazu kommen, dass *Vermögen* in manchen Pfaden einen hohen Beitrag leistet und in anderen als Feature vollständig verschwindet. Das bedeutet, dass diese Variable lokal innerhalb des Netzes unterschiedlich stark gewichtet wird. In diesem Sinne ist es auch nicht plausibel, von *einer* Variable *Vermögen* zu sprechen, da sich diese Variable gewissermaßen durch das Netz vervielfältigt und abhängig von der lokalen Umgebung im Netz einen unterschiedlichen Beitrag zur Weiterverarbeitung und letztlich zum Output leistet. Die einzelnen Variablen werden im Laufe des Trainings irreversibel miteinander assoziiert. Jedes Profil aktiviert gewissermaßen ein eigenes lokales Subnetz (Berghoff, Neu & von Twickel, 2020; BSI, 2022, 2024a; Goodfellow, Bengio & Courville, 2016).

Komplexität des Datensatz Wenn wir die Modellkomplexität einmal ausblenden und den Blick auf den Datensatz wenden, dann wird Folgendes deutlich. Der *German Credit Data* Datensatz umfasst 20 Attribute, bestehend aus kategorialen und numerischen Features. Die kategorialen Features besitzen mehrere Ausprägungen (z.B. *Sparkonto* mit 4 Klassen, *Beschäftigungsdauer* mit 5 Klassen). Wir können nun eine Approximation des gesamten Kontextraums U vornehmen.

Hierbei nehmen wir der Einfachheithalber an:

- 13 kategoriale Features mit durchschnittlich 4 Ausprägungen $\Rightarrow 4^{13} = 67.108.864$ Kombinationen und
- 7 numerische Features, diskretisiert in je 100 mögliche Werte $\Rightarrow 100^7 = 100.000.000.000.000$ Kombinationen.

$$U \approx 4^{13} \cdot 100^7 = 67.108.864 \cdot 100.000.000.000.000$$

Nehmen wir weiterhin an, wir könnten jede mögliche Eingabe in 1 ms prüfen:

$$U \approx 4^{13} \cdot 100^7 \,\text{ms} \approx 212.800.811.770.67 \,\text{Jahre}.$$

Dies verdeutlicht, dass die vollständige testbasierte bzw. empirische Überprüfung von *Hinreichendheit* und *Notwendigkeit* bereits bei einem relativ kleinen Datensatz wie *German Credit*, der ausschließlich in einem Laborumfeld für Forschungszwecke genutzt wird, *de facto* unrealistisch ist. Einschränkend ist noch hinzuzufügen, dass wir bei einer Sicherheitstestung Parameter setzen können, um die Menge der möglichen Kontexte weiter auf die Menge der realistischen einzugrenzen. Insofern würde man eine kontrollierte Stichprobe $u \in \mathcal{U}$ ziehen. Abhängig von den Eigenschaften der Daten, des Modells und der Anwendung lässt sich der Testraum somit enorm beschränken, eine vollständige Testung in allen relevanten Anwendungskontexten bleibt nichtsdestotrotz unrealistisch. Die entscheidende Beobachtung ist, dass wir nicht alle *rechnerisch* möglichen Welten in der Testung berücksichtigen können. Ab einer bestimmten Komplexität besteht immer die Möglichkeit, dass eine unbekannte Inputkonstellation zu einem nicht gewünschten Modellverhalten und Modelloutput führt, was eine potenzielle Bedrohung für die informationelle Selbstbestimmung darstellen kann (BSI, 2022, 2024a; Mothilal & Tan, 2021).

Implikationen Aus der *Komplexität der Modelle* und der *Komplexität des Datensatzes* folgen der Verlust der globalen Koeffizienten und die Unabhängigkeit der Variablen und damit zwei wichtige Einschränkungen für die Güte von Erklärungen, welche mit der Notwendigkeit und Suffizienz von Merkmalen arbeiten.

1. Ein Merkmal oder auch eine Merkmalskombination kann in einem Kontext eine oder mehrere der Kriterien der *Idealen Erklärung* erfüllen und in einem anderen nicht. Zum Beispiel kann die Variable *Vermögen* in einem Kontext als notwendig und/oder hinreichend für den Output identifiziert werden, während dies in einem anderen Kontext nicht gilt.

2. Die Gefahr, dass eine vermeintlich robuste Erklärung doch falsifiziert wird, lässt sich aufgrund der komputationalen Restriktionen nicht vermeiden, da eine vollständige Testung auch nur aller realistischer Kontexte aus U unter bestehenden Beschränkungen in komplexen Fällen nicht realistisch ist (Liang et al., 2022).

Partielle Erklärung für Modellausgaben Konsequenterweise benötigen wir eine Flexibilisierung, welche das Modell der idealisierten Erklärung auf eine Stichprobe $u \in \mathcal{U}$ eingrenzt, die partielle Erklärung. (Mothilal & Tan, 2021) schlagen in ihrem Ansatz zwei

Metriken vor, welche für uns interessant sind, die α und β -Metriken. Die Ausgangslage ist die oben skizzierte Komplexität:

Die α Metrik misst, wie häufig ein bestimmtes Feature-Set in u tatsächlich **kausal notwendig** für den Output y^* ist:

$$\alpha = \Pr(\mathbf{x}_j = a \text{ ist Ursache von } y^* \mid \mathbf{x}_j = a, y = y^*)$$

In anderen Worten, wie häufig können wir tatsächlich messen, dass die Bedingung $\mathbf{x}_j = a$ die Bedingungen für Minimalkausalität erfüllen. Intuitiv lassen sich hohe Werte für α so interpretieren, dass $x_j = a$ häufig notwendig ist, während niedrige Werte für α auf eine tendenziell zu vernachlässigende Korrelation hinweisen.

Gegeben: $\mathbf{x}_j = a$ und $y = y^*$ (tatsächliche Beobachtung)

- Wenn $\alpha \approx 1$, dann ist $x_j = a$ *fast immer notwendig*, um $y = y^*$ zu erzeugen.
- Wenn $\alpha \approx 0$, dann ist $x_j = a$ *selten notwendig*, um $y = y^*$ zu erzeugen.

Die β – Suffizienz wiederum misst, ob die Manipulation von $\mathbf{x}_j = a$ in u hinreichend ist, um y^* zu erzeugen. Dies wird getestet, indem $\mathbf{x}_j = a$ konstant gehalten wird und alle anderen Features variiert werden.

$$\beta = \Pr(y = y^* \mid \mathbf{x}_j \leftarrow a)$$

In anderen Worten, wie häufig können wir tatsächlich messen, dass die Bedingung $\mathbf{x}_j = a$ die Bedingungen für Suffizienz erfüllt. Um Suffizienz nachzuweisen, müsste $\mathbf{x}_j = a$ in möglichst vielen Kontexten unabhängig von allen anderen Faktoren zu $y = y^*$ führen.

Nehmen wir modellhaft einmal an: $\mathbf{x}_j = a$ und $y = y^*$ (tatsächliche Beobachtung)

- Wenn $\beta \approx 1$, dann ist $x_j = a$ *fast immer hinreichend*, um $y = y^*$ zu erzeugen.
- Wenn $\beta \approx 0$, dann ist $x_j = a$ *fast nie hinreichend*, um $y = y^*$ zu erzeugen.

Beide Fälle sind im Kontext des sogenannten Deep Learning mit tiefen neuronalen Netzen und hoch-dimensionalen Datensätzen ohne das aktive manipulieren der Hyperparameter sehr unwahrscheinlich.

Die Intuition hinter den beiden Metriken lässt sich wie folgt zusammenfassen. Wenn ich die Features der Stichprobe $\mathbf{u} \in \mathcal{U}$ fixiere und nur die Bedingung $\mathbf{x}_j = a$ geändert wird, dann fasst dies das Konzept der Notwendigkeit. Wenn ich hingegen alle Features von $\mathbf{u} \in \mathcal{U}$ frei variieren lasse und nur $\mathbf{x}_j = a$ konstant halte, dann messe ich damit die Suffizienz von $\mathbf{x}_j = a$.

Probleme Metriken und epistemisches (Rest-)Risiko Die *alpha* und *beta*− Metrik können womöglich eine bestimmtes Maß an Interpretierbarkeit für die Stichprobe $\mathbf{u} \in \mathcal{U}$ herstellen. Das Problem an dieser Stelle ist wie beschrieben (6.2), dass die Menge der möglichen Transformationen mit der Anzahl der Parameter und der Dimensionalität der Trainingsdaten exponentiell anwächst. Mit wachsender Modellkomplexität steigt die Wahrscheinlichkeit, dass selbst große Stichproben aus U nicht repräsentativ sind.

Wenn wir diese Argumentation mit den Funden über die Komplexität (6.3) und deren Sicherheitsrisiko (6.4) zusammentragen, können wir für einen irreduziblen Mangel an Interpretierbarkeit und Sicherheit argumentieren, welcher sich aufgrund dieser Komplexität nicht vollständig durch Testung einfangen lässt. Die besprochenen Verfahren bleiben bis zu einem gewissen Grad kontextabhängig und probabilistisch und sind daher nicht vollständig verlässlich generalisierbar. Wir können konsequent von einem *irreduziblen (epistemischen) Risiko* sprechen.

7.4 Die Herstellung von Sicherheit als Verstehen

Resümieren wir nun die wichtigsten Desiderata für die Gütekriterien, die im nächsten Abschnitt dann genauer spezifiziert werden.

7.4.1 Desiderata Anwendung der Gütekriterien

Im Rahmen der beiden Klassen von Gütekriterien (7.1) sollte dieser Herausforderung wesentlich mit einem dualen Prozess begegnet werden, bei dem technische und institutionelle Gütekriterien als Komponenten eines holistischen Wechselverhältnis zu verstehen sind (für

eine detaillierte Beschreibung siehe den Ablaufplan (7.4.2)):

1. *Institutionelle Gütekriterien:* Diese Komponente umfasst das holistische Risikomanagement zu dem insbesondere der oben skizzierte Trialog gehört, als ein Forum zur Herstellung von Transparenz, menschlicher Aufsicht und Rechenschaftspflicht. In diesem Rahmen müssen die normativen Anforderungen, sowie technischen Möglichkeiten und Grenzen transparent gemacht werden. Auch die theoretischen (bzw. ontologischen) Aspekte von Interpretierbarkeit und Komplexität, wie oben dargelegt, könnten in diesem Forum Verwendung finden. Im Zentrum dieser Komponente steht der transparente Trialog über die Möglichkeiten, sowie Grenzen der technischen Produzierbarkeit von *epistemischer Sicherheit* zur Förderung menschlicher Autonomie und Autorenschaft.

2. *Technische Gütekriterien:* Diese Komponente umfasst die formale und programmbasierte Modellierung, Testung und Evaluation, während der Entwicklung und Anwendung der Systeme. Dabei könnten unter anderem die oben entwickelte Bedingungsontologie, sowie die besprochenen technischen Kriterien und Metriken zur Anwendung kommen. Spezifischer umfasst diese das Konstruieren einer möglichst robusten Erklärung unter Berücksichtigung ihrer Unsicherheit. Dies dient dem Ziel *a.)* eine möglichst hohe Kohärenz der Erklärung herzustellen und *b.)* Informationen über das bestehende epistemische Restrisikos zu gewinnen.

3. *Holismus:* Beide Komponenten müssen in einem produktiven Dialog stehen, das heißt, dass sie sich während des gesamten Lebenszyklus eines KI-Systems immer anhand der am Einzelfall gewonnenen Erkenntnisse neu informieren müssen. Dies kann entlang der technischen Achse dazu dienen, neue Methoden zu testen oder bestehende zu verbessern und entlang der institutionellen, den Dialog mit von KI Output betroffenen, natürlichen Personen besser zu informieren.

4. *Fallbasiert und sachlogisch:* Jeder Fall bzw. jeder Kontext generiert einen eigenen Möglichkeitsraum für die Sicherheit der Betroffenen. Entsprechend können bzw. sollten wahrscheinlich auch keine eindeutigen und fallagnostischen Schwellenwerte *a priori* angegeben werden, die determinieren, dass der Prozess Sicherheit als Verstehen gelingt bzw. scheitert. Es können konsequenterweise nur Gelingensbedingungen, das

heißt Gütekriterien, formuliert werden, aber kein Algorithmus. Doch dass der Prozess gelingt, kann zumindest wahrscheinlicher werden.

7.4.2 Ablaufplan Anwendung Gütekriterien

Basierend auf den Ergebnissen bis hierhin können wir nun einen Ablaufplan für die Anwendung der Gütekriterien vorstellen. Dieser folgt einem methodischen Reduktionismus dahingehend, dass wir diesen nur auf die hier beschriebenen Parameter eingrenzen. Das KI bezogene Risikomanagement umfasst, wie oben aus der Gesetzgebung abgeleitet, noch deutlich mehr, aber wir legen hier den Fokus auf das *Sicherheitsmanagement im Hinblick auf Interpretierbarkeit*. Auch werden viele weitere praktisch relevante Aspekte ausgelassen, wie zum Beispiel Audits oder Mitarbeiter:innen-Schulungen. Auch die Umsetzung bestehender Rechtsnormen wie der Datenschutz-Grundverordnung (DSGVO) oder dem Bundesdatenschutzgesetz (BDSG) ist hier bereits vorausgesetzt.

Methodisch und sprachlich ist der Ablaufplans bereits an dem BSI Grundschutzkompendium orientiert. In Orientierung an dem Grundschutzkompendium ist zudem von konkreten Institutionen (Softwarefirma, Kreditinstitut, Krankenhaus und so weiter) abstrahiert worden, um zunächst einen allgemeinen Ablaufplan zu erhalten, der dann anhand spezifischer Institutionen und Hochrisiko Anwendungen instanziiert werden kann. Auch die ersten Schritte sind in den Ablauf zur Einrichtung eines Informationssicherheitsmanagementsystems (ISMS) nach den BSI Standards 200-1 bis 200-3 integriert. Dies erlaubt es Forschenden, Betreibenden und Akteur:innen der Informationssicherheit sich besser zu orientieren. In Referenz auf die BSI-Standards wird nur der jeweilige Schritt genannt, zum Beispiel Verantwortung durch die Leitungsebene, aber auf eine Beschreibung ist hier zu verzichten.[71] Der Übersicht halber sind die Rollenbeschreibungen hier häufig auf den XAI-Beauftragten reduziert. Damit ist nur markiert, dass der jeweilige Schritt in der Hauptverantwortung des Beauftragten liegt, aber auch andere Akteur:innen, insbesondere Entwickelnde, die Leitungsebene oder das ISMS-Team für diesen Schritt relevant sein könnten. Darüber hinaus sind alle genannten Begriffe, Methoden und Konzepte in diesem Ablaufplan so zu verstehen, wie sie in *dieser Arbeit* entwickelt wurden, womit Unterschiede

[71] Für eine anfängerfreundlichen Einblick in das Grundschutzkompendium zum Aufbau eines ISMS siehe IT-Grundschutzschulung: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/IT-Grundschutz/Zertifizierte-Informationssicherheit/IT-Grundschutzschulung/it-grundschutzschulung_node.html

mit anderen Quellen möglich sind.

Die Tabelle stellt den Ablauf der Anwendung schemenhaft als Schritte oder auch Bausteine dar. Unter (7.5.13) ist die Abarbeitung dieser Schritte am Fallbeispiel German Credit beschrieben. In der Tabelle ist in der Beschreibung immer das jeweilige Kapitel, aus dem dieser Aspekte gewonnen wurde referenziert:

Tabelle 2: Ablaufplan Anwendung Gütekriterien Hochrisiko-KI-Systeme

Nr.	**Titel**	**Beschreibung**
1	Ausgangslage	Eine Institution nutzt für die Ausführung ihrer Geschäftsprozesse den Modelloutput von KI-Systemen, welche auf (tiefen) künstlichen neuronalen Netzen basieren. Beispiele für eine solche Institution könnten ein Kreditinstitut, eine medizinische Forschungseinrichtung oder eine Sicherheitsbehörde sein. (3.1), (6.2), (9.1)
2	Risikoklassifikation nach EU KI-VO	Der Output dieser Modelle (Klassifikation, Prädikation oder synthetische Inhalte) bedingt die Zuweisung von grundrechtsrelevanten Ressourcen zu natürlichen Personen (zum Beispiel Kredite, Abschlüsse, Medikamente). Folglich wird das Modell als Hochrisikomodell eingestuft. Diese Klassifikation muss bereits eingeleitet werden, wenn die Absicht formuliert wurde, ein entsprechendes System einzusetzen und ggf. während des Prozesses der Entwicklung und Inbetriebnahme korrigiert und protokolliert werden. (3.2), (3.3)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	Titel	Beschreibung
3	Verantwortung durch die Leitungsebene	Die Leitungsebene ist in der Verantwortung in Kenntnis zu sein, über mögliche Hochrisiko-Anwendungen in der eigenen Institution und den damit einhergehenden Risiken. Die Leitungsebene ist damit in der Pflicht, diese Anwendung beim Initiieren eines Informationssicherheitsmanagementsystems (ISMS) und beim Ausarbeiten der Informationssicherheitsrichtlinie zu berücksichtigen. Nach ISO-27001/BSI-Grundschutz initiiert die Leitungsebene die Einrichtung eines ISMS.(7.2.6)
4	Informationssicherheitsrichtlinie	Neben den anderen Bestimmungen, die gemäß Grundschutzkompendium Teil der Richtlinie werden sollten, könnte Sicherheit als Verstehen und Vertrauen inklusive der Förderung menschlicher Autonomie in Bezug auf Hochrisiko-KI-Systeme explizit als Ziel für das ISMS ausformuliert werden. (4), (5)
5	ISMS	Als Teil des Informationssicherheitsmanagementsystems (ISMS) richtet das ISMS-Team die Stelle einer XAI-Beauftragten ein. Aufgrund der Komplexität der Systeme sollte die ernannte Person umfassende technische, ethische und rechtliche Kompetenzen in Bezug auf datenverarbeitende Systeme mitbringen. (7.2),(7.3)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	**Titel**	**Beschreibung**
6	Holistisches Risikomanagement	Einführung eines iterativen Systems mit Betriebsanleitung, Folgeabschätzung und menschlicher Aufsicht. Die Implementierung und kontinuierliches Monitoring ist Aufgabe der XAI-Beauftragten, welche wiederum von dem ISMS-Team überwacht wird. Die XAI Beauftragte sollte in diesem Kontext glaubhaft gegenüber allen Beteiligten darlegen, wie ein kontinuierliches, iteratives Risikomanagement im Falle der fraglichen Anwendung realisiert wird. (7.2.1), (7.2.6), (7.5.8), (7.5.10)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	Titel	Beschreibung
7	Betriebsanleitung	Die XAI-Beauftragte richtet eine Betriebsanleitung ein. In Bezug auf Interpretierbarkeit umfasst diese folgende Punkte: • Zweckbestimmung des Modells und Beschreibung des Algorithmus, des Datensatzes und weitere technische Eigenschaften. (7.2.1) • Beschreibung der Risiken. Dies umfasst sowohl reale, gemessene Risiken, als auch mögliche Risiken, die aus den Modelleigenschaften resultieren. (6.2) • Erläuterung der normativen Zwecke von Interpretierbarkeit. (4),(5.9) • Erläuterung der analytischen Dimensionen von Interpretierbarkeit. (5.9), (7.2.3), (7.2.4) • Erläuterung der Methoden zur Herstellung von Interpretierbarkeit. (7.5.9), (7.5.10) • Erläuterung der Risiken und Grenzen dieser Methoden. (6.2), (6.4.1), (7.5.14) • Erläuterung der Methoden zur Herstellung von Rechenschaftspflicht, menschliche Aufsicht und Transparenz (folgende Schritte). (7.2.2)
8	Protokoll	Die XAI-Beauftragte richtet ein ständiges Protokoll als Teil des kontinuierlichen, iterativen RM ein. Hier werden sämtliche Auffälligkeiten, Anomalien und Fehlverhalten festgehalten. Es empfiehlt sich, nach der Anonymisierung personenbezogener Daten, dieses Protokoll zu veröffentlichen bzw. potenziell Betroffenen des Systems zugänglich zu machen. (7.2.1)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	**Titel**	**Beschreibung**
9	Forum Sichere KI	Einrichtung eines institutionellen Forums als Kommunikationsschnittstelle zwischen Modell, Betroffenen und Stakeholdern. Durch dieses Forum wird der Prozess der Reziprozität und der Rechenschaftspflicht und Transparenz institutionalisiert. (7.2.2),(7.2.3)
10	Informationspflicht und Einwilligung	Alle Personen, die von dem Modelloutput betroffen sind oder sein könnten, werden im Vorfeld umfassend darüber informiert. Hierzu kann die Betriebsanleitung oder eine didaktisch aufbereitete Zusammenfassung der Betriebsanleitung dienen. Die Information umfasst mindestens die folgenden Aspekte: • Zweck und der Umfang der Modellprognose. • Information über die (Hyper-)Komplexität von bestimmten KI-Modellen und das daraus resultierende Risiko. (6) • Die Möglichkeiten und prinzipiellen Grenzen von Erklärungen. (6.4.1) • Die Möglichkeit über das Verfahren im Rahmen des Trialog Forums aufgeklärt zu werden, inklusive über die zur Anwendung kommenden Metriken und das epistemische (Rest-)Risiko. (7.2.2) • Die Möglichkeit zur menschlichen Intervention in die resultierende Entscheidungskette. (4), (7.2.1) Basierend auf dieser Informationslage muss sich die XAI-Beauftragte eine informierte Einwilligung der potenziell betroffenen Personen einholen, dass das Modell wie dargestellt genutzt werden darf.

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	Titel	Beschreibung
11	Kohärenz und Kontinuität	Die XAI-Beauftragte muss die Robustheit bzw. Sensitivität des Modells überprüfen. Hierfür kann der gesamte Methodenkorpus für sichere und vertrauenswürdige KI-Systeme angewandt werden. So können Stichproben aus der anvisierten Grundgesamtheit gezogen und mit adversialen Samples eine Testung stattfinden. (7.5.8) Für Details sind einschlägige Überblickarbeiten zu empfehlen: (Guo et al., 2025; Hu et al., 2024; Huang et al., 2020)
12	Menschliche Aufsicht und Korrigierbarkeit	Bevor ein Modelloutput Konsequenzen für eine natürliche Person S hat, muss eine *menschliche Intervention möglich* gewesen sein. Bevor das System zur Anwendung kommt, stellt die XAI-Beauftragte sicher, dass der Output des KI-Systems nicht unmittelbar irreversible Konsequenzen für S hat. Falls sich aufgrund der Art und Weise, die das Modell in den Geschäftsablauf integriert ist, eine solche Irreversibilität ergibt, ist unbedingt von der Verwendung abzusehen. Unter bestimmten Bedingungen wäre dies dann auch illegal. (4), (7.2.1)
13	Anwendungsbetrieb und menschliche Aufsicht	Das Modell kommt im ständigen Geschäftsbetrieb zum Einsatz, die Prognosen werden genutzt. Die XAI-Beauftragte überwacht die Nutzung. Dabei protokolliert sie etwaige Probleme und interveniert, wenn ein Fehlverhalten des Modells auftaucht. (7.2.1), (7.2.3)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	**Titel**	**Beschreibung**
14	Fallsituation	Natürliche Person S_x, die von dem Output eines entsprechendes Modells betroffen ist, wird in das Forum geladen, um gemeinsam mit der XAI-Beauftragten und weiteren Stakeholdern zu ermitteln, ob eine autonomiefördernde Erklärung konstruiert werden kann. (7.2.2), (7.2.5), (6.4.1)
15	Initiierung Erklärungsmodell und Protokoll	Die XAI-Beauftragte initiiert den Prozess der Konstruktion eines Erklärungsmodells. Das Ziel ist die Ermittlung einer lokalen Erklärung des Outputs für Person S_x. Es werden alle Schritte des folgenden Vorgehens protokolliert und später der Betriebsanleitung als Anhang bzw. dem ständigen Sicherheitsprotokoll für Hochrisikosysteme angefügt. (7.2.1), (7.2.2)
16	Transparenz	Die XAI-Beauftragte klärt S_x über das Vorgehen auf, über das Vorgehensprotokoll, die Möglichkeiten und prinzipiellen Grenzen von Erklärungen und erläutert erneut die zur Anwendung kommenden Metriken und das epistemische (Rest-)Risiko. (7.5.14)
17	Auswahl der Features	XAI-Beauftragte und S_x bestimmen gemeinsam die Eingabemerkmale, deren Einfluss erklärt werden soll. (7.5.7)
18	Berechnung der Notwendigkeit und Suffizienz	Für die gewählten Merkmale werden die Notwendigkeits- (α) und Suffizienzmetrik (β) berechnet. (7.5.10), (7.3.4)
19	Konstruktion der Erklärung und Explikation des epistemischen Risikos	Es wird schriftlich eine Erklärung modelliert. Hierzu gehört eine vorläufige Beantwortung der Fragen entlang der *epistemologischen Parameter* (7.2.4) und der *kontextuellen Sachlogik* (7.2.5) durch die XAI-Beauftragte.

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	**Titel**	**Beschreibung**
19a	Epistemische Parameter	• Kann S_x die Bedingungen, die zum Modelloutput führten, nachvollziehen? • Kann S_x die Entscheidungslogik des Modells basierend auf der im Rahmen der Einwilligung zur Verfügung gestellten Informationen nachvollziehen? • Kann S_x dargestellt werden, welche Faktoren hätten anders sein müssen, um einen anderen Output zu erlangen? (7.2.4), (7.5.13)
19b	Kontextuelle Sachlogik	1. *Narrative Plausibilität:* • Ist die hohe/niedrige Suffizienz der angezielten Variable sachlogisch überzeugend? • Ist es sachlogisch überzeugend, dass die angezielte Variable notwendig/nicht notwendig für den Output ist? 2. *Kontrafaktische Autonomie:* • Kann der Person transparent gemacht werden, unter welchen kontrafaktischen Szenarien sie einen anderen Modelloutput hätte erreichen können? • Sind diese Szenarien sachlogisch fair/unfair? (7.2.5)

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	Titel	Beschreibung
20	Trialog und Entscheidung	Im Forum wird S_x das Erklärungsmodell dargestellt. Es wird dargelegt, inwieweit der Modelloutput auf interpretierbare Features zurückzuführen ist. S_x wird auch über die Grenzen der algorithmischen Erklärbarkeit, inklusive vorhersehbarer als auch unvorhersehbarer Risiken informiert. (7.5.13) Nach etwaigen Rückfragen wird gemeinsam entschieden, ob das Restrisiko tragbar ist und das Modell zur Anwendung kommen kann. (6.4.1) Wenn wir im normativen Rahmen dieser Arbeit dieses Moment auf eine binäre Entscheidungssituation reduzieren wollten, dann lautet die in diesem Prozess final zu beantwortende Frage: *Erweitert die Verwendung des Modelloutputs die Autonomie und Autorinnenschaft des betroffenen Subjekts S_x oder schränkt sie diese ein?* (4) Wenn in diesem Verfahren keine Einigung erzielt werden kann, bleibt S_x die Möglichkeit den Gerichtsweg zu gehen. Hier müsste ermittelt werden, ob die verantwortliche Instanz ihrer gesetzlichen Pflicht angemessen nachgekommen ist. (7.2.1)
21	Protokollierung und Verbesserung	Der Verlauf und die Ergebnisse werden protokolliert. Etwaige sicherheitsrelevante Erkenntnisse zum Beispiel über Risiken werden dem Protokoll hinzugefügt. Verbesserungsvorschläge und Ansätze zur Innovation des Risikomanagements werden an das ISMS-Team und ggf. an die Leitungsebene herangetragen (7.2.1), (7.2.6), (7.5.15).

Fortsetzung auf nächster Seite

Fortsetzung von vorheriger Seite

Nr.	Titel	Beschreibung
22	Konsequenzen	Aus dieser Entscheidung können unterschiedliche Konsequenzen resultieren. In Bezug auf den fraglichen Fall könnte ein anderer, transparenter Algorithmus (Entscheidungsbaum, lineare Regression) als Alternative herangezogen werden. In Bezug auf das Modell könnte auch der Entschluss gefasst werden, dass das Risiko generell zu groß ist und es aus dem Betrieb entfernt wird oder es wird sich entschieden, das Modell nochmal in die Entwicklungsphase zu schicken. (7.5.15), (8.1)

7.5 Programmbasierte Evaluation

Ziel: Die Schritte *14-22* aus dem Ablaufplan (7.4.2) adressieren die praktische Umsetzung dieser Gütekriterien. Die kleine Fallstudie dient primär dem Zweck die Umsetzung der technischen Kriterien anhand eines vereinfachten Fallbeispiels zu *simulieren*. Dadurch werden die bis hierhin etwas abstrakten Gütekriterien für die geneigten Leser:innen anschaulicher und wir können darauf aufbauend Implikationen, Möglichkeiten zur Anwendung und Erweiterungen, sowie Grenzen diskutieren. Im einem realen Fall würden Teile der hier beschrieben Ergebnisse anschließend nochmal in das Informationssicherheitskonzepts, der Betriebsanleitung und das Sicherheitsprotokolls überführt werden. Um Redundanz zu vermeiden, wird auf die Ausformulierung dieser Dokumente verzichtet.

Disclaimer Die Evaluation dient der Veranschaulichung und Problematisierung und *nicht* dem Zweck eine vollständige und umfassende Validierung dieser Gütekriterien vorzunehmen, dies ließe sich nur anhand einer Serie von Fallstudien mit deutlich komplexeren Modellen und Datensätzen realisieren (siehe auch (8.2.1)). Daraus folgt auch insbesondere, dass das hier veranschlagte Modell, als auch die generierten Erklärungen *nicht* für den direkten Praxistransfer geeignet sind. Ich ziele darauf ab, den Grundaufbau plausibel zu *simulieren* und behaupte nicht, dass die generiere Erklärungslogik in realen Fällen überzeugt. Um die

kontextuelle Sachlogik für Kreditanalysen vollständig zu erfassen, wäre eine noch deutlich ausführlichere Evaluation und Auswertung nötig, dies würde jedoch den Rahmen dieser Arbeit sprengen.

7.5.1 Quellcode

Für die kleine Evaluation wurde Folgendes programmiert:

- *German Credit Data lesbar*: Decodierung des Datensatzes in ein lesbares Format zur Veranschaulichung für den Lesenden
- *German Credit ANN*: Standardarchitetktur zur Prognose von Kreditrisiko
- *Selection*: Auswahlprogramm, welches 10 diverse Profile aus dem Datensatz zieht
- *Evaluation*: Berechnung der α- und β-Metrik für die 20 Variablen der 10 Profile
- *Evaluation_ details*: Wiederholung der Evaluation mit detailliertem Report über die genauen Werte der Variablen für die beiden Metriken (Optional, nicht notwendig für die Analyse)

Für den Quellcode, sowie die Ergebnisdateien und weitere technischer Details siehe das Githup Repositorium (Niehus, 2025).

7.5.2 Vorgehen

Für die Evaluation gehen wir wie folgt vor. Zunächst müssen ein paar Grundannahmen getroffen werden (7.5.3). Dann wird das Ausgangsszenario der Fallstudie kurz vorgestellt (7.5.4). Dann wird kurz das Modell (7.5.5) und der Datensatz und die Auswahlkriterien der zu prüfenden Profile beschrieben (7.5.6), um darauf aufbauend die Auswahlkriterien der Profile und Variablen zu bestimmten (7.5.7) Anschließend geht es um den Transfer des hier entwickelten Ansatzes auf geeignete XAI-Methoden (7.5.8) und welche Dimensionen von XAI-Methoden für die Evaluation geeignet (7.5.9) sind und eine Methodenauswahl begründet (7.5.10). Dann wird der genaue Ablauf der Anwendung dieser Methoden im Rahmen des Ablaufplans für dieses Fallbeispiel beschrieben (7.5.12) und die Ergebnisse vorgestellt (7.5.13). Dann wird auf das Restrisikos (7.5.14) hingewiesen und die Ergebnisse im Rahmen des Forums (7.5.15) eingebettet. Den Implikationen (8.1), sowie die Diskussion

(8.2) wird sich gesondert in dem anschließenden Hauptkapitel Sicherheit als epistemisches Vertrauen (8) gewidmet. Wichtig ist noch der Hinweis, dass alle Schritte nur ein mögliches Vorgehen illustrieren und in der Realität durchaus andere Wege denkbar sind.

7.5.3 Grundannahmen und Restriktionen

Da diese Gütekriterien sich nur mit dem Thema Interpretierbarkeit und verwandten Aspekten wie Transparenz und menschliche Aufsicht beschäftigen, gilt es eine Reihe von Grundannahmen zu treffen, welche als Teil des Risikomanagements bereits realisiert sein müssen, damit die Gütekriterien zur Anwendung kommen können:

- Die Infrastruktur für das Risikomanagement ist vorhanden und eingerichtet. Das bedeutet die technischen, organisatorischen und personellen Ressourcen zur Realisierung des Ablaufplans (7.4.2) sind gegeben. Dies umfasst auch die Ressourcen für die Risikoidentifikation und Risikominimierung, Testung und Monitoring.

- Die Verantwortlichen haben eine umfassende Daten Gouvernance sichergestellt. Dies umfasst unter anderem die Prüfung auf Vollständigkeit und Repräsentativität der Daten, das Sicherstellen der Vertraulichkeit und Integrität, sowie das Einhalten aller gesetzlichen Bestimmungen.

- Die Ressourcen und Fähigkeiten für das Erstellen eines technischen Reports sind vorhanden, inklusive einer detaillierten Beschreibung der technischen Details des Systems, Performance-Metriken und des Monitorings.

- Die Ressourcen für die Reports über fehlerhaftes und/oder unerwünschtes Verhalten des Modells sind vorhanden und einsetzbar.

7.5.4 Fallstudie

Wir können uns nun eine idealisierte Verantwortliche, die XAI-Expertin imaginieren, die im Verbund mit anderen Akteur:innen die Aufgabe hat, das umfassende Risikomanagement und die Konformitätsbewertung von Hochrisiko Systemen zu realisieren. Diese könnte nun den theoretischen und angewandten Teil dieser Arbeit nutzen, um den Teil eines auf Hochrisiko-KI-Systeme spezifiziertes ISMS aufzubauen, der für die Blackbox-problematik bzw. der Sicherheit als Interpretierbarkeit zuständig ist. Dabei kann sich die

XAI-Beauftragte für einige Bausteine des Risikomanagements inspirieren lassen oder gar den Ablaufplan zu weiten Teilen wie angegeben, übernehmen. Die verantwortliche Person könnte eine ähnliche Evaluation durchführen wie unten, deren Resultate dann Teil der Betriebsanleitung und des Protokolls werden sollten, anhand dessen dann ggf. weitere Maßnahmen ergriffenen werden müssten.

Es ist wie oben dargelegt essenziell für die Sicherheit von KI-Systemen, dass jedes Modell individuell evaluiert wird. Folgende Frage kann als leitgebend für die XAI-Beauftragte gelten:

Leitfrage 1: *Welche Methodenkonstellation als Teil meines Risikomanagementsystems testet die fragliche Anwendung auf die rechts-ethischen Standards dieser Gütekriterien?*

Für die Evaluation nehmen wir einen Fall heraus, welcher rechtlich und ethisch kompatibel ist mit den oben skizzierten epistemischen Zustand des (nicht-)Verstehens. Das heißt es geht um folgendes Szenario. Eine natürliche Person, das epistemische Subjekt S_x, ist von dem Output eines Hochrisiko-KI-Systems betroffen. S_x hat nach geltender Rechtslage (DSGVO (71) und EU KI-Verordnung Art. 86) ein Recht auf Erklärbarkeit. Wie dieser zunächst vage Anspruch *ex-ante* als Zustand des Verstehens im hier beschriebenen Sinne realisiert werden kann, soll hier diskutiert werden.

7.5.5 Modelleigenschaften

Gegenstand der Evaluation ist die Anwendung und Diskussion der Gütekriterien, nicht die Interpretierbarkeitsgüte des hier verwendeten Modells oder die Steigerung der Performance. Daher sollte die Komplexität der Basismodelle so kontrollierbar wie möglich gehalten werden. Entsprechend nutzen wir hier die Standardarchitektur mit folgenden Eigenschaften. Die Eingabeschicht nimmt das gesamte Profil mit 62 One-Hot-Codierten Werten entgegen. Die Eingabe wird dann durch zwei versteckte Schichten mit 6 und 8 Neuronen und ReLU-Aktivierungsfunktion geschickt. In der letzten Schicht wird der Output mittels Sigmoid in eine kontinuierliche Wahrscheinlichkeitsverteilung zwischen 0 und 1 transformiert. Das Netzwerk wird mit randomisierten Gewichten initialisiert (Niehus, 2025).

7.5.6 Datensatz

Der German Credit Data Datensatz ist ein Standarddatensatz aus dem maschinellen Lernen, der häufig für Studien im Bereich Explainable AI genutzt wird (bspw. (Mothilal & Tan, 2021)). Er ist öffentlich zugänglich unter der UCI-Lizenz: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data. Der Datensatz enthält Kreditprofile und umfasst 1000 Instanzen. Die Profile sind gelabelt mit gutem oder schlechtem Kreditrisiko. Er umfasst 700 gute und 300 schlechte Kredite. Die Profile umfassen 20 Variablen, 7 numerische und 13 kategoriale Variablen. Die Variablen beschreiben sozioökonomische Eigenschaften der Profile, zum Beispiel Kredithöhe, Beziehungsstatus oder Anzahl der Beschäftigungsjahre (Details unter (Niehus, 2025)).[72]

7.5.7 Profile und Features

Auswahlkriterien der Profile Es werden 5 Profile mit dem Label negativ und 5 Profile mit dem Label positiv aus dem Datensatz gezogen. Es werden positive und negative Fallbeispiele ausgewählt, da ein Verlust der informationellen Selbstbestimmung auch im Falle einer positiven Kreditentscheidung vorliegen kann. Für die kritische Frage nach der informationellen Selbstbestimmung in einem gegebenen Fall spielt der Output zunächst eine untergeordnete Rolle, obzwar dieser natürliche erhebliche Konsequenzen für die Betroffenen hat. Der Algorithmus *Selection* stellt mittels eines Max-Min-Auswahlverfahrens sicher, dass sich die Profile in ihren wesentlichen Merkmalen stark unterscheiden (Niehus, 2025).

Auswahl der zu untersuchenden Features Es wurde bewusst ein für aktuelle Verhältnisse relativ kleiner Datensatz gewählt, mit nur 20 Features. Das erlaubt es eine Evaluation aller 20 Features vorzunehmen. Dabei wird die Evaluation in einem zweischrittigen Prozess vollzogen (Niehus, 2025).

1. Im ersten Schritt werden nur die Features betrachtet, die im engeren Sinne als ökonomische zu verstehen sind. Ökonomische Features sind hier definiert als solche, die direkt messbare monetäre Werte darstellen, wie zum Beispiel Einkommen, Kredithöhe oder Wertpapiere, als auch indirekte Indizes für monetäre Liquidität wie der Status des Girokontos oder die Beschäftigungsdauer. Nach dieser Definition lassen sich die

[72] Der Datensatz ist im Original auf Englisch verfügbar, ich habe diesen mittels gängiger KI-Werkzeuge wie *DeepL* und *Gemini* auf deutsch übersetzt.

folgenden 12 Variablen als ökonomische definieren:

- **Kreditbetrag** – absolute Höhe des aufgenommenen Kredits
- **Kreditgeschichte** – bisheriges Zahlungsverhalten
- **Kreditverwendungszweck** – Zweck des beantragten Kredits
- **Dauer in Monaten** – Laufzeit des Kredits
- **Ratenhöhe** – monatlich zu zahlende Kreditrate
- **Sparkonto / Wertpapiere** – vorhandene Rücklagen oder Sicherheiten
- **Vermögen** – zusätzliche Vermögenswerte
- **Status des Girokontos** – Kontoführung, Überziehungsmöglichkeiten
- **Beschäftigt seit** – Dauer der aktuellen Beschäftigung als Einkommensindikator
- **Andere Ratenverpflichtungen** – weitere laufende finanzielle Verpflichtungen
- **Weitere Bürgen / Schuldner** – Vorhandensein von Mitkreditnehmenden oder Bürgen
- **Anzahl bestehender Kredite** – bereits laufende Kredite

2. Im zweiten Schritt werden die verbleibenden 8 Variablen und damit der sozio-demographische Kontext mit einberechnet. Damit kann der Einfluss der weiteren Lebensumstände wie der Beziehungsstatus, das Alter oder die Wohnsituation auf den Modelloutput berechnet werden:

   - **Alter** – Lebensalter der Antragstellenden
   - **Familienstand / Geschlecht** – Familienstand und Geschlechtsangabe
   - **Wohnsitzdauer** – Dauer des aktuellen Wohnsitzes
   - **Wohnsituation** – Miet- oder Eigentumssituation
   - **Unterhaltspflichtige Personen** – Anzahl der zu versorgenden Personen

- **Telefon** – Vorhandensein eines Festnetzanschlusses
- **Ausländischer Arbeiter** – Staatsangehörigkeit (binäre Kodierung)
- **Beruf** – ausgeübte berufliche Tätigkeit

Diese Eingrenzung hat einen entscheidenden Vorzug. Die ökonomischen Variablen sollten der Logik einer Kreditanalyse entsprechend die dominanten sein. Daher lässt sich auch durch eine isolierte Betrachtung dieser mitunter feststellen, ob diese schon hinreichend sind, um eine starke Erklärung zu konstruieren und falls nicht, ggf. auf unerwünschtes, unter Umständen diskriminierendes Modellverhalten hinweisen. Weiterhin können wir etwas simplifiziert, aber für diese Zwecke ausreichend auf die Kontingenz ökonomischer Features hinweisen, das heißt das Subjekt S_x hätte prinzipiell die Möglichkeit, diese Variablen zu ändern. Demgegenüber sind viele demographische Merkmale, wie das biologische Alter, Herkunft oder Geschlecht nicht oder nicht ohne erheblichen Aufwand zu ändern, was auch eine der Gründe dafür ist, warum diese aus ethischen Gründen in einer Kreditanalyse keine relevante Rolle spielen sollten.[73]

7.5.8 Transfer der Gütekriterien auf geeignete XAI-Methoden

Nachdem die XAI-Beauftragte in diesem Szenario die Features ausgewählt hat, stehen ihr für die Testung und Evaluation der XAI-Methodenkorpus zur Verfügung. Die Beauftragte muss sich gewissermaßen fragen, welche Dimensionen und Methoden den im letzten Kapitel beschriebenen Desiderata gerecht werden, um den Ablaufplan (7.4.2) für einen fraglichen Fall zu realisieren. In diesem Abschnitt werden in Kürze die wichtigsten Dimensionen von XAI-Methoden abgeschritten und einige Punkte diskutiert, die die jeweilige Dimension für die Methodenauswahl der XAI-Beauftragten interessant machen könnte.[74]

7.5.9 Dimensionen und Methoden

In (Pouyanfar et al., 2018) werden drei relevante Dimensionen von XAI-Methoden unterschieden:[75]

[73] Bei einer detaillierten Analyse sozio-demographischer Merkmale wäre eine genauere Taxonomie vonnöten. Diese binäre Zuweisung (ökonomisch=kontingent und demographisch=notwendig) hält einer genaueren Betrachtung natürlich nicht stand, kann aber als Näherung für diese Evaluation hilfreich sein.

[74] Wobei dies natürlich keine detaillierte Analyse aller XAI-Methoden substituiert, sondern nur einige für diese Arbeit relevante Aspekte hervorhebt.

[75] Vom Autor auf deutsch übersetzt und paraphrasiert, J.N.

1. **Dimension — Passive vs. Aktive Ansätze**

   - **Passiv**: Nachträgliche Erklärung trainierter neuronaler Netzwerke
   - **Aktiv**: Aktive Veränderung der Hyperparameter (Netzwerkarchitektur, Trainingsprozesses, Verlustfunktion und so weiter) zur besseren Interpretierbarkeit

2. **Dimension — Typen von Erklärungen** Um den Output (Prognose/Klassifikation) zu erklären:

   - **Beispiele**: Bereitstellung von Beispiel(en), die als ähnlich oder als Prototyp(en) gelten können
   - **Attribution**: Zuweisung von Verdienst (oder Schuld) zu Eingabemerkmalen (zum Beispiel Merkmalsbedeutung, Salienzmasken)
   - **Verborgene Semantik**: Beiträge bestimmter verborgener Neuronen/Schichten zum Modelloutput decodieren
   - **Regeln**: Extraktion logischer Regeln (z.B. Entscheidungsbäume, Regelmengen und andere Regel-Formate)

3. **Dimension 3 — Lokale vs. Globale Interpretierbarkeit**

   - **Lokal**: Erklärung der Vorhersagen des Netzwerks für einzelne Fälle (zum Beispiel eine Salienzmaske für ein Eingabebild)
   - **Semi-lokal**: Mittelstelle zwischen lokal und global. Beispielsweise aggregierte Erklärung einer Gruppe ähnlicher/vergleichbarer Eingaben
   - **Global**: Erklärung des Netzwerks als Ganzes (zum Beispiel ein Regelsatz/ein Entscheidungsbaum)

Dimension 1: Entlang der ersten Dimension ist festzustellen, dass sich der methodische Ansatz dieser Gütekriterien auf passive Erklärungen bezieht. Entlang dieser Dimension möchte ich eine wichtige Unterscheidung einführen. Die passiven Methoden, wie sie in

dieser Arbeit vorgestellt werden, zielen darauf ab, *post-hoc* das *externe Verhalten* des Modells zu modellieren. Dieser Ansatz hat eine erwähnenswerte (und ebenso problematische) Verwandtschaft zum Behaviorismus. Es geht tendenziell darum, dass drittpersonal beobachtbare und messbare Verhalten dieser Modelle statistisch zu modellieren. Dem gegenüber steht, genau wie beim klassischen Behaviorismus, die Ausklammerung der *internen* Repräsentationen und die sich aus diesen bildenden intrinsische Entscheidungslogik des Modells (Graham, 2023; Z. C. Lipton, 2016). In realen Umgebungen scheinen aktive Ansätze, zum Beispiel die Modifikation der Verlustfunktion oder der Modell Architektur oftmals unrealistisch. Auch die Gesetzgebung sieht zum jetzigen Zeitpunkt keine Einschränkung auf aktive Ansätze vor. Wobei es durchaus denkbar wäre, dass der Gesetzgeber bzw. die zuständigen Behörden Leitlinien und Whitepaper rausgibt, um die Interpretierbarkeit *by design* mit entsprechend aktiven Ansätzen für Training oder Modellarchitektur zu erhöhen (European Commission, 2024a; O. Müller & Lazar, 2024). Daher sollten wir perspektivisch das Potenzial von aktiven Ansätzen auch nicht unterschätzen (Di Marino et al., 2025; Zhang et al., 2021). Der Fakt, dass diese in dieser Arbeit bis hierhin ausgeklammert wurden, liegt erstens darin begründet, dass der hier vorgeschlagene Ansatz zunächst anhand der abstrakt-allgemeinen Eigenschaften von einfachen künstlichen neuronalen Netzen diskutiert werden soll. Zweitens erlauben die begrenzten zeitlichen Ressourcen zum Abfassen dieser Arbeit keine umfassende Integration dieser Ansätze.

Dimension 2: Entlang der zweiten Dimension ist festzuhalten, dass für die ersten drei Typen von Erklärungen (Beispiele, Attribution und interne Repräsentationen) ermittelt werden müsste, inwieweit sie als Instanzen der beiden vorgestellten Metriken gelesen werden können. Dies wird beispielhaft im nächsten Kapitel vorgenommen (7.5.10). Im Allgemeinen ist darauf hinzuweisen, dass sie besonders effektiv genutzt können, wenn sie tatsächlich prototypische Beispiele oder rezeptive Felder innerhalb von Netzwerken identifizieren, die interpretierbaren Konzepten korrespondieren, um damit einen gewissen Grad an Interpretierbarkeit zu ermöglichen (Zhang et al., 2021). Der vierte Typ sind die regelbasierten Ansätze. Dabei ist festzuhalten, dass der in den letzten Kapiteln entwickelte theoretische und formale Ansatz im weiteren Sinne den regelbasierten Ansätzen zuzuordnen ist bzw. diese ergänzt und präzisiert. Die Bedingungsontologie stellt ein Gütekriterium für die regelbasierten Ansätze auf. Dabei muss die Erklärung letztlich nicht in einer formalen Sprache transkribiert werden, methodisch kann es aber durchaus sinnvoll sein, etwa für

die XAI-Beauftragte, die Erklärung, zum Beispiel unter Zuhilfenahme der Modallogik und Prädikatenlogik zu formalisieren. Letztlich muss dem betroffenen Subjekt S_x die Erklärung aber in natürlicher Sprache dargelegt werden. Daraus folgt allerdings nicht, dass die Modelle intrinsisch logisch interpretierbar sein müssten. Stattdessen müssen die Verfahren, die zum Einsatz kommen, die verborgene Semantik so weit entschlüsseln, dass sie in entsprechende Regeln integriert werden kann (siehe auch (7.5.14)) (Zhang et al., 2021). Wobei der Ausdruck verborgene Semantik meines Erachtens auch etwas irreführend sein kann. Semantik wird nämlich oftmals als Bedeutungslehre definiert, die die Relation zwischen Zeichen (Signifikant) und Bezeichnetem (Signifikat) betrifft. Dabei ist die Semantik als Bedeutungslehre oftmals auf die für Menschen versteh- und interpretierbaren, lebensweltliche Signifikate bezogen. In diesem engeren Sinne erlernen die Modelle keine wirkliche Semantik, wie oben kurz skizziert wurde. Es ist ja gerade der kritische Sachverhalt, dass die sie oftmals keine versteh- und interpretierbare Semantik ausprägen, auch keine verborgene.

Dimension 3: Entlang der dritten Dimension können wir argumentieren, dass lokale und globale Interpretierbarkeit in diesem Kontext als komplementär zu betrachten sind. Die globale Interpretierbarkeit kann als Teil der Testung wichtig sein, zum Beispiel auf Robustheit (Meng et al., 2022), (Goodfellow, Bengio & Courville, 2016, S. 240f.). Hier kommen in der Regel statistisch gemittelte Werte über Datensamples zum Einsatz, die uns Auskunft über die Modell- und Datenqualität geben können. Diese Parameter können auch indirekt über diskriminierende Muster aufklären, durch Hinweise auf Proxy-Bias oder unbalancierte Datensätze. Sie ist in den hier behandelten Fällen (7.5) stärker mit dem Teil des Risikomanagements vor der Inbetriebnahme und Verbreitung des Models assoziiert. Die lokale Interpretierbarkeit hingegen ist besonders relevant, wenn es darum geht, ein betroffenes Subjekt in einem fraglichen Falle, eine Autonomie-fördernde Erklärung an die Hand zu geben. Hier ist es der kritische Einzelfall, der entscheidend ist.

7.5.10 Anwendung auf XAI-Methoden

In der Forschung ist ausgearbeitet worden, dass bestimmte Methoden aus dem XAI-Korpus als Annäherungen der *alpha*- und *beta*-Metriken gelesen werden können (Halpern, 2016; Mothilal & Tan, 2021). Dies zeigt auf, inwieweit der analytische Rahmen in die bestehende

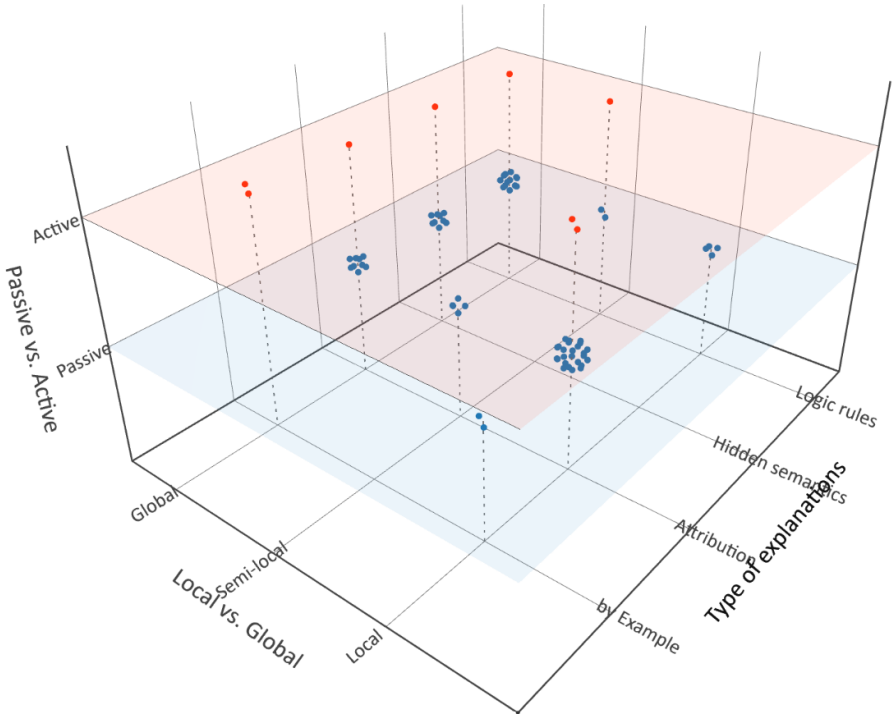

Abbildung 8: Dimensionen XAI-Methoden

Quelle: Darstellung nach (Pouyanfar et al., 2018)

XAI-Landschaft integriert werden kann. Die Begründung soll hier in Kürze für zwei interessante Methoden reproduziert werden. Zusätzlich möchte ich die zugrundeliegende Intuition dahinter kurz am Beispiel Kreditanalyse aufzeigen. Nehmen wir wieder das abstrakte neuronales Netz $f : \mathbb{R}^d \to \mathbb{R}$ mit dem Input $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ als Eingabevektor und $y = f(\mathbf{x})$ als Modelloutput. Für eine konkrete Eingabeinstanz $\mathbf{x}_0 \in \mathbb{R}^d$ ergibt sich der Modelloutput $y^* = f(\mathbf{x}_0)$. Für einen spezifischen Fall $f(\mathbf{x}_0)$ besteht die Aufgabe darin, diesen Output unter den vorangestellten Desiderata zu erklären.

Kontrafaktische Methoden und die *alpha*-Metrik Kontrafaktische Erklärungen nach (Wachter, Mittelstadt & Russell, 2018) suchen nach einer minimalen Modifikation von $\mathbf{x}_0$, die nötig ist, um den Modelloutput zu $y = \neg y^*$ zu ändern.

Formal ausgedrückt wird die folgende Funktion optimiert:

$$\arg\min_{x'} \max_{\lambda} \left[\lambda \cdot (f_w(x') - y')^2 + d(x_i, x')\right]$$

Dabei werden zwei Optimierungsfunktionen durchgeführt. Erstens soll die Änderung an x, x', die zum veränderten Modelloutput y' führt minimiert werden. Gleichzeitig soll der

Gewichtungsparameter λ iterativ erhöht werden, um $f_w(x')$ nah an y' zu bringen. Die Parameter des Modells f_w müssen bei der Suche nach x' gleich gehalten werden, damit sich die eigentlichen Modelleigenschaften nicht verändern. Der Regularisierungsterm $d(x_i, x')$ bestraft bei der Optimierungssuche zu große Abweichungen vom Original. Im Ergebnis wird ein Kompromiss aus Fehlerterm und Änderung von x angestrebt.

Intuition Es ist leicht zu sehen, dass die kontrafaktische Optimierungsfunktion als Instanz der Alpha-Metrik gelesen werden kann. Dies kann am Beispiel der Kreditanalyse für das betroffene Subjekt S illustriert werden:

> *Welche minimalen Änderungen hätte S machen müssen, damit der Algorithmus eine andere Entscheidung trifft?*

Damit entspricht die Optimierungsfunktion einer Suche nach der Teilmenge der Merkmalswerte ($\mathbf{x}_j \leftarrow a'$ und $a' \neq a$), die notwendig ist, um den Modelloutput zu verändern ($y \neq y^*$). In anderen Worten, welche minimale Änderung in dem Kreditbewerbungsantrag hätte gereicht, um den Output zu ändern.

LIME und die *beta*-Metrik Die attributionsbasierte Methode LIME (Local Interpretable Model-agnostic Explanations) nach (Ribeiro, Singh & Guestrin, 2016) versucht mittels eines weniger komplexen, interpretierbaren Modells das Verhalten der Blackbox nachzuahmen, indem es die Blackbox *lokal* approximiert (Ribeiro, Singh & Guestrin, 2016).

Formal ausgedrückt wird die folgende Funktion optimiert:

$$\xi(\mathbf{x}_0) = \arg\min_{g \in \mathcal{G}} [L(f, g, \pi_{\mathbf{x}_0}) + \Omega(g)]$$

Diese Formel beinhaltet folgende Bestandteile:

- $\xi(\mathbf{x}_0)$: Die *lokale Erklärung* für $\mathbf{x}_0$. Diese besteht zum Beispiel aus einer linearen Regression, die die Entscheidung von f in der Umgebung von $\mathbf{x}_0$ nachbildet.
- $g \in \mathcal{G}$: Menge aller zugelassenen Erklärungsmodelle, z. B. lineare Regressionen mit sparsamen Gewichten. LIME ist bewusst auf einfache Modelle beschränkt, um

Interpretierbarkeit zu ermöglichen.

- $L(f, g, \pi_{\mathbf{x}_0})$: Der *Treueverlust*. Dieser Term misst, wie gut das einfache Modell g das Verhalten von f für Punkte in der Nachbarschaft von $\mathbf{x}_0$ approximiert. Formal:

$$L(f, g, \pi_{\mathbf{x}_0}) = \sum_{\mathbf{z} \in Z} \pi_{\mathbf{x}_0}(\mathbf{z}) \cdot (f(\mathbf{z}) - g(\mathbf{z}'))^2$$

  Dabei ist:

  - Z: die Menge von artifiziell erzeugten Störpunkten $\mathbf{z}$ in der Umgebung von $\mathbf{x}_0$
  - $f(\mathbf{z})$: Output des Blackbox-Modells für Punkt $\mathbf{z}$
  - $g(\mathbf{z}')$: Output des Erklärungsmodells für die interpretierbare Repräsentation $\mathbf{z}'$
  - $\pi_{\mathbf{x}_0}(\mathbf{z})$: eine Gewichtungsfunktion (z. B. ein Gauß-Kernel), welche die Punkte nahe an $\mathbf{x}_0$ höher gewichtet

- $\Omega(g)$: Eine *Komplexitätsstrafe* für das Modell g. Dieser Regularisierungsterm bevorzugt einfachere Erklärungen, etwa durch:

  - Anzahl aktiver Merkmale (ℓ_0-Norm bei linearen Modellen)
  - Tiefe oder Anzahl von Regeln bei Entscheidungsbäumen oder Regelmodellen

Zusammengefasst sucht LIME also nach einem Modell g, das das Verhalten von f *lokal* gut imitiert, aber zugleich möglichst einfach bleibt.

Intuition Funktional vollzieht LIME etwas vergleichbares wie die *beta*−Metrik (Mothilal & Tan, 2021). Das lineare Surrogatmodell beantwortet gewissermaßen folgende Frage:

Wie stabil verhält sich der Algorithmus, bei Variationen des Eingabeprofils $\mathbf{x}_0$*?*

Die Intuition dahinter können wir veranschaulichen, wenn wir für $\mathbf{x}_j = a$ beispielsweise das Einkommen nehmen. LIME ermittelt nicht direkt die Suffizienz von $\mathbf{x}_j = a$, sondern approximiert sie über die *Sensitivität* des Modelloutputs gegenüber lokalen Störungen. Diese Sensitivität lässt sich dann aus den Gewichten des (linearen) Surrogatmodells ablesen. Wenn die Gewichte hoch sind, dann ist $\mathbf{x}_j = a$ wahrscheinlich nicht hinreichend für den

Modelloutput und wenn sie niedrig sind, dann lässt dies auf einen hohen Beitrag von $\mathbf{x}_j = a$ schließen. Nehmen wir zur Veranschaulichung die beiden Extremfälle an. Wenn sich trotz starker Variation der Umgebung der Modelloutput nicht ändert, dann spricht dies für eine hohe Suffizienz der Variable Einkommen. Dies liegt darin begründet, dass die Variable Einkommen *alleine* den Output stabil halten kann. Wenn andererseits bei nur leichter Variation der Umgebung das Modell bereits seinen Output wechselt, dann spricht dies für einen niedrigen Grad an Suffizienz des Einkommens, da ein fixes Einkommen bereits bei leichten Änderungen des Profils den Modelloutput nicht stabil halten kann. Das Surrogatmodell bildet gewissermaßen die Sensibilität bzw. Stabilität des fraglichen Features, zum Beispiel Einkommen ab. Die beobachtete *Sensibilität bzw. Stabilität* ist die lokale Suffizienz (Mothilal & Tan, 2021; Ribeiro, Singh & Guestrin, 2016).

7.5.11 Methodenauswahl und technische Details

Perturbationsmethode Die Analyse der vorgestellten Dimensionen und Methoden aus dem Feld XAI veranschaulicht, inwieweit der hier vertretene Ansatz in das weitere Feld der Explainable AI integriert werden kann. Dabei ist es nicht zwingend, die genannten Methoden zu nutzen, die XAI-Beauftragte muss sich nur für eine gut begründete Auswahl entscheiden. Da es hier nicht um die Evaluation bestimmter XAI-Methoden geht, sondern um die Simulation und Diskussion der Gütekriterien, habe ich keine Surrogatmodelle wie zum Beispiel LIME oder SHAP gewählt (Mothilal & Tan, 2021). Stattdessen habe ich mich hier in Konsistenz mit dem oben entwickelten Interpretierbarkeitsbegriff für eine Perturbationsmethode entschieden. Das heißt, es werden *direkt* die Eingabevariablen der Profile variiert, um zu messen, ob der Output stabil bleibt bzw. kippt. Damit können wir überprüfen, ob und unter welchen Bedingungen sich der Modelloutput ändert (Notwendigkeit, α) bzw. stabil bleibt (Suffizienz, β) (Ivanovs, Kadikis & Ozols, 2021; Watson et al., 2021).

Technische Details Für *kategoriale Variablen* wird die α-Metrik berechnet, indem alle möglichen Kategorien der Variable einmal getestet werden. Wenn die Variable im Original mit *One-Hot* als $(0, 0, 0, 1)$ kodiert war, werden alle alternativen Kodierungen $(1, 0, 0, 0)$, $(0, 1, 0, 0)$ und $(0, 0, 1, 0)$ jeweils ausprobiert. Für *numerische Variablen* wird zunächst ein plausibler Wertebereich definiert, der durch die 5- und 95-Perzentile der

Trainingsverteilung begrenzt ist. Innerhalb dieses Bereichs werden neun äquidistant verteilte Werte getestet. Wenn mehrere Werte das Modellurteil verändern, wird derjenige mit dem geringsten Abstand zum Originalwert gespeichert. Sobald eine dieser Änderungen den Output kippt, wird der Test abgebrochen und der kontrafaktische Wert (die Kategorie oder der Zahlenwert) gespeichert. Die β-Metrik funktioniert für beide Variablentypen ähnlich. Es werden 2 000 zufällige, leicht variierte Profile erzeugt. Die zu untersuchende Variable wird dabei in allen Profilen auf denselben (fixierten) Wert gesetzt, während alle anderen Merkmale zufällig variiert werden. Anschließend wird der Anteil der Profile berechnet, bei denen das Modellurteil stabil bleibt. Liegt dieser Anteil über der gewählten Schwelle von 0,70 % gilt die Variable in diesem Setting als *hinreichend*. Die Wahl dieser Schwelle ist recht willkürlich, ein Aspekt der im Abschnitt (8.1) nochmal aufgegriffen wird. Entsprechend der Funktionen werden die Ergebnisse der α-Metrik binär (α=1 oder α=0) dargestellt und die Ergebnisse der β-Metrik als integer zwischen 0 und 1 (z.B. β=0.83) (Niehus, 2025).

7.5.12 Anwendung

Für die Simulation der Schritte ist hier nun folgendes Schema angewandt worden:

1. Das Programm *Selection* zieht 10 möglichst diverse Profile aus dem Originalen Datensatz.

2. Wir wählen die 12 ökonomischen Features aus. Zur Vergleichbarkeit werden für alle 10 Profile die gleichen Features betrachtet.

3. Es wird die α - Metrik berechnet.

4. Es wird die β - Metrik berechnet.

5. Erklärung konstruieren und Ergebnisse festhalten.

6. Schritt 2. bis 5. werden für die verbleibenden 8 soziodemographischen Merkmale wiederholt.

7. Fortfahren mit Schritt 19 im Ablaufplan.

7.5.13 Externes Modellverhalten

Nach der programmbasierten Berechnung der Metriken (Schritt 18) kommt der XAI-Beauftragten nun die Aufgabe zu, diese zu interpretieren und als Grundlage in die zu konstruierende Erklärung einfließen zu lassen (Schritt 19). Sehen wir uns eine aggregierte Darstellung der 10 Profile innerhalb der epistemologischen Parameter (7.2.4) und der kontextuellen Sachlogik (7.2.5) an (Niehus, 2025).

Ökonomische Variablen

Kontextuelle Sachlogik. Narrative (Schein-)Plausibilität An einem Ende des Spektrums stehen die Profile mit hohen Suffizienz-werten für die einzelnen Variablen. Für die Profile **1, 4, 8, 9** und **10** ließen sich auf den ersten Eindruck Erklärungen mit einer gewissen narrativen und sachlogischen Plausibilität generieren. In diesen Fällen liegen für die ökonomischen Variablen (z. B. Kreditbetrag, Vermögen, Dauer in Monaten, Sparkonto/Wertpapiere) stabile β-Werte über $0,70$ vor. Dies erlaubt Erklärungen, die für S_x sachlogisch plausibel erscheinen mögen, etwa dass ein höheres Vermögen oder ein geringerer Kreditbetrag die Wahrscheinlichkeit einer positiven Entscheidung erhöht. Doch bei genauerer Betrachtung fällt auf, dass bei *allen* genannten Profilen es nicht einzelne ökonomische Faktoren sind, die in vielen Fällen hinreichend sind, sondern *alle* ökonomischen Variablen sind in oftmals deutlich über 70 % der Fälle hinreichend für den Output. Wenn jede einzelne Variable schon hinreicht, um die Modellprognose zu erklären, dann kann genau genommen von Erklärung keine Rede sein. Das Modell ist deutlich überdeterminiert. Wenn S_x die berechtigte Frage nach den Bedingungen für die Modellprognose stellt, kann jede Variable als bestimmender Faktor genannt werden. Dies ist eher ein narrativer Freifahrtschein, beliebig plausible Geschichten zu konstruieren. Auf der anderen Seite des Spektrums stehen jene Profile mit sehr niedrigen Suffizienz-werten. Für die Profile **2, 3, 5, 6** und **7** liegen die Werte oftmals deutlich unter 30 %. Entsprechend wäre es nicht möglich, den Betroffenen plausibel darzulegen, welche Faktoren das externe Modellverhalten beeinflusst haben. Hier liegt das andere Extrem vor, mittels der Metriken lässt sich überhaupt keine Erklärung konstruieren. Die externe Entscheidungslogik zeigt in Bezug auf die β-Metrik eine starke Tendenz zu den Extremen. Das heißt in 50 % der Fälle sind (fast) alle ökonomischen Features hinreichend und in 50 % der Fälle ist (fast) keine

Variable hinreichend, um den Modelloutput zu erklären (Niehus, 2025).

Kontrafaktische Autonomie Bei den Profilen **1, 4** und **5** lässt sich eine gewisse kontrafaktische Autonomie mit sachlogischer Plausibilität herstellen. Es kann plausibel dargelegt werden, dass etwa Vermögen oder Kreditbetrag notwendig zum Modelloutput beigetragen haben. Diese Merkmale sind prinzipiell durch menschliches Handeln oder ökonomische Entscheidungen beeinflussbar. Für S_x besteht somit die Möglichkeit, eine gewisse rationale Vorstellung davon zu entwickeln, welche Handlungen in einem realistischen Rahmen den Modelloutput hätten beeinflussen können. Auf der anderen Seite zeigt sich auch bei diesen Profilen eine gewisse Übersensitivität für kontrafaktische Bedingungen. Das heißt bei den genannten Profilen ist schon die Mehrheit der Variablen kontrafaktisch relevant. Besonders anschaulich lässt sich dieses Problem bei Profil **6** studieren. In diesem Fall reicht eine Änderung von jeder einzelnen Variable aus, um den Modelloutput zu kippen, während zugleich keine Variable hinreichend für den Output ist. Es ist sachlogisch nicht sehr plausibel, dass die Modifikation jeder einzelnen Variable bereits die Prognose kippt. Dies spricht für eine Übersensitivität des Modells auf nur leichte Änderungen der Profildaten. Weiterhin kann für die Profile **2, 3, 7** und **8** nur bedingt oder überhaupt keine kontrafaktische Autonomie im Sinne einer handlungsrelevanten Erklärung hergestellt werden. Die α-Werte zeigen keine stabilen Bedingungen, die Rückschlüsse auf beeinflussbare Faktoren des Modelloutputs erlauben. Der Zusammenhang zwischen Eingabevariablen und Modelloutput ist mitunter so schwach, dass S_x weder nachvollziehen noch hypothetisch beeinflussen kann, wie eine alternative Entscheidung zustande gekommen wäre. Somit bleiben ökonomische Parameter (z. B. Kreditbetrag, Laufzeit, Vermögen) ohne klar rekonstruierbares kausales Verhältnis zur Entscheidung (Niehus, 2025).

Soziodemographische Variablen Erweitern wir das Bild um eine aggregierte Darstellung der soziodemographischen Variablen.

Kontextuelle Sachlogik. **β-Metrik (Suffizienz):** Bei den Profilen **1, 4, 8, 9** und **10** wiederholt sich zunächst das Problem der Überdeterminiertheit. Es sind fast alle soziodemographischen Merkmale hinreichend für den Output. Hinzu kommt, dass eine solche Kausalstruktur deutlich auf sachlogische bzw. normativ problematische Muster hinweist: Alter, Familienstand, Geschlecht oder Wohnsituation sind mitunter bereits

hinreichend, um den Modelloutput stabil zu halten. Solche Erklärungen wären in zweierlei Hinsicht entmündigend. Erstens handelt es sich um mitunter diskriminierende Muster und zweitens kann bei dieser Entscheidungslogik nicht einmal überzeugend dargelegt werden, welche dieser Merkmale entscheidend gewesen sind, da sie alle hohe Suffizienz-Werte aufweisen. Auf der anderen Seite ist bei den Profilen **2, 3, 5, 6** und **7** keine der soziodemographischen Variablen hinreichend. Da diese Merkmale idealerweise keinen bzw. nur einen marginalen Beitrag leisten sollten, ist dies sachlogisch plausibel (Niehus, 2025).

α-Metrik (Notwendigkeit): Bei den Profilen **1, 4, 5, 6** und **10** führt bereits die Modifikation einzelner demographischer Variablen dazu, dass der Modelloutput kippt. Diese kontrafaktische Sensitivität kann auf diskriminierende Muster hindeuten: Das Modell reagiert in zentralen Entscheidungspfaden auf nicht beeinflussbare persönliche Eigenschaften (z. B. Alter, Geschlecht, familiäre Situation). In solchen Fällen wäre die Modellentscheidung problematisch, weil die (nahezu) Notwendigkeit dieser Merkmale bedeutet, dass der Output ohne sie nicht stabil bleibt. Für die Profile **2, 3, 7, 8** und **9** zeigen die α-Werte hingegen keine signifikanten Muster, demographische Variablen üben dort keinen dominanten kontrafaktischen Einfluss aus (Niehus, 2025).

Zusammenfassung Aus dieser kurzen Analyse können für die technischen Reports die folgenden Ausprägungen des *externen Modellverhaltens* unterschieden werden:

1. Sachlogische Scheinplausibilität: Eine Entscheidungslogik die eine gewisse Plausibilität bereitstellt, doch keine ermächtigende Kausalstruktur offenbart. Gründe hierfür können die Überdeterminiertheit oder die Übersensitivität des Modells sein.

2. Sachlogisch problematische Entscheidungslogik: Eine Entscheidungslogik die unsinnige und/oder diskriminierende Muster aufweist.

3. Kontrafaktischen Entmündigend: Es lassen sich keine sinnvollen kontrafaktischen Handlungsszenarien aufzeigen. Entweder stellt die generierte Erklärungslogik keine kontrafaktisch relevanten Faktoren zur Verfügung oder diese hätten nicht sinnvollerweise von S_x geändert werden können.

4. Blackbox: Es lässt sich überhaupt keine Erklärung generieren. Selbst die externe Entscheidungslogik bleibt vollkommen intransparent.

Abhängig davon, welches Modellverhalten sich bei der Evaluation auftut, könnten dann diese Ausprägungen ergänzt bzw. differenziert werden. Die vier Ausprägungen können in einigen Fällen in Kombination auftreten. Wie gesehen kann das externe Modellverhalten sowohl sachlogische *und/oder* kontrafaktisch entmündigend sein *und/oder* problematische Entscheidungslogiken enthüllen. Die Ausprägung Blackbox tritt alleine auf, da sie das Szenario beschreibt, indem überhaupt keine Entscheidungslogik ermittelt werden kann.

Partielle epistemische Autonomie: Die vier Ausprägungen stellen die Kontrastfolie bereit, um die Voraussetzungen für *partielles epistemisches Vertrauen* zu formulieren:

- Hoch narrative sachlogische Plausibilität der Entscheidungslogik
- Keine Hinweise auf unsinnige und/oder problematische Entscheidungslogik
- Sachlogisch plausible, das heißt autonomiefördernde, kontrafaktische Szenarien

In einem solchen Fall wäre aus der Perspektive der Evaluation nicht mehr und nicht weniger als die notwendigen Voraussetzungen für die Möglichkeit von Verstehen und Vertrauen geschaffen. Diese Ergebnisse gilt es dann in den weiteren Prozess der Anwendung der Gütekriterien zu integrieren.

7.5.14 Epistemisches Restrisiko

Weiterhin gilt es im Forum das epistemische Restrisiko aufzuklären (Schritt 20). Wenn wir relativ hohe Werte bei den beiden zur Hand genommen Metriken und eine hohe sachlogische Plausibilität aufweisen können, bleibt das epistemische Restrisiko nichtsdestotrotz enorm. Die XAI-Beauftragte muss transparent machen, dass sie mittels beider Metriken eine Approximation des dritt-personal messbaren *externen* Modellverhaltens vornimmt. Es handelt sich dabei nicht notwendigerweise um ein akkurates Modell der *internen* Modelllogik. Im besten Falle erlaubt diese Approximation eine *partielle* logische Autonomie, das heißt S_x hat eine partielle Sicherheit, welche Faktoren entscheidend waren für die Prognose oder anders hätten sein müssen, um die Prognose zu verändern. Das heißt beispielhaft, dass S_x den Modelloutput hätte verändern können, indem die Variable Einkommen erhöht wird, was nicht heißt, dass das isolierte Feature Einkommen tatsächlich als solches in der internen Modelllogik repräsentiert wird und dass es zudem die Relevanz für die interne Logik des Modells hat, die wir ihm zuschreiben. Insbesondere sind Proxy-Strukturen nicht

auszuschließen, die zu einem diskriminierendem oder unsinnigem Modellverhalten führen können. Das Modell könnte Alter als Proxy für Erwerbsbiografien nutzen, ohne dass dieses Feature selbst inhaltlich entscheidend wäre. In anderen Worten selbst ein *extern* plausibles Modellverhalten kann *intern* auf fragilen oder diskriminierenden Artefakten beruhen.

7.5.15 Forum, Dokumentation und Verbesserungen

Als Teil des iterativen Risikomanagements werden die Ergebnisse der Evaluation, insbesondere die aggregierten Ergebnisse und die Auffälligkeiten bei Einzelprofilen protokolliert. In einem realen Fall würden die Ergebnisse für den Einzelfall S_x der Person im Forum präsentiert und diskutiert werden. Die wichtigsten Entwicklungen und Erkenntnisse werden protokolliert. Hier ist es vor allem wichtig, Entwicklungen festzuhalten, die sich nicht antizipieren ließen, das heißt zum Beispiel Auffälligkeiten bei der Testung oder unerwartete Rückfragen von S_x. Im idealen Fall einer hohen Erklärungsgüte des externen Modellverhaltens stellt die Erklärung die Voraussetzung für den weiteren trialogischen Prozess im Forum dar. Das bedeutet insbesondere, dass die eigentliche Erklärungsgüte nicht mehr und nicht weniger ist, als ein notwendige aber keinesfalls hinreichende Bedingung für Sicherheit. Die Ergebnisse werden abschließend in einem Report dem ISMS-Team und unter Umständen der Leitungsebene präsentiert. Es wird über die Weiterverwendung bzw. ggf. Modifikation des Modells entschieden.

8 Sicherheit als epistemisches Vertrauen

8.1 Implikationen

In diesem Abschnitt sollen in Kürze einige wichtige Beobachtungen, die aus der Beschreibung der Ergebnisse resultieren, dargelegt werden.

8.1.1 Normativität der Metriken und Rechenschaftspflicht

Als erste Beobachtung ist festzuhalten, dass die Evaluation noch einmal die Relevanz eines trialogischen Forums unterstreicht. Die Metriken sprechen selbst mit einem geeigneten Interface nicht für sich selbst. Dies ist zum Beispiel anhand der Feature Auswahl, als auch der Schwellenwerte für die Metriken zu erkennen. Die verantwortlichen Personen müssen unter Umständen im Trialog mit S_x entscheiden, nach welcher Methode die Samples für die

Evaluation gezogen werden sollen und wie groß das Sample sein soll. Auch muss entschieden werden, wie hoch der Anteil des Samples sein muss, der den Output stabil hält, um von Suffizienz zu sprechen. In diesem Falle haben wir uns für $beta = 0.7$ entschieden. Es ist eine Entscheidung, dass von 2000 Featurekombinationen, 70 % ausreichen sollen, um von Suffizienz zu sprechen. Diese und andere methodische und konzeptionelle Entscheidungen und ihre statistische Natur werden dann später durch ein die Semantik von Ausdrücken wie Notwendigkeit, Suffizienz und kausale Logik verschleiert. Doch dies sollte nicht darüber hinwegtäuschen, dass all diese Entscheidungen in einem Möglichkeitsraum infinit vieler anderer methodologischer Optionen stattfinden, für die sich die Verantwortlichen auch hätten entscheiden können. Dass sich letztlich so und nicht anders entschieden wurde, ist eine normative Festlegung, die vorschreibt, dass ein spezifischer Schwellenwert x ausreichend sein *soll*. Dieser Sachverhalt, die Entscheidungen und die Begründungen für diese sollten S_x im Forum transparent dargelegt werden. Unter Umständen ist zu erwägen, inwieweit S_x an diesen Entscheidungen zu beteiligen ist.

8.1.2 Recht und Behörden

An dieser Stelle kommt dem Gesetzgeber und den zuständigen Behörden auch eine verstärkte Rolle zu. In erster Linie ist der EU KI-Verordnung ein Defizit in Bezug auf die Unterbestimmtheit ihrer Vorschriften zu attestieren. Eine Implikation, die sich sicherlich politisch und justiziell noch dahingehend rechtfertigen lässt, dass es eine Vielfalt an Methoden und Verfahren gibt, zum Beispiel der hier vorgestellte Ansatz, welcher konsistent mit der Gesetzgebung ist. Auch mag das Argument eines paternalistischen Staates überzeugen, der Verbraucher:innen, Unternehmen und Wissenschaft nicht exakte Methoden und Metriken vorschreiben sollte, der ihre lokalen und individuellen Fähigkeiten und Ressourcen ignoriert. Aber spätestens an dieser Stelle sollten die zuständigen Behörden in die Pflicht genommen werden. Es bedarf international einheitlicher Leit- und Richtlinien, die, wenn sie nicht verpflichtend sind, als Orientierung für die Praxis eines ISMS in Bezug auf Hochrisiko-KI-Systeme zur Anwendung kommen sollten. Diese Arbeit hat einige Parameter für eine legale, aber auch legitime Anwendung von Hochrisiko-KI-Systemen skizziert. Diese Parameter sollten von den zuständigen Behörden noch deutlich präzisiert werden, damit Verantwortliche wie Entwickelnde oder Betreibende eine gewisse Sicherheit bekommen, innerhalb welcher technischen und institutionellen Parametern sie ein solches System

einsetzen können. Dies schafft Rechtssicherheit und Vertrauen in die Gesetzgebung und die Verwaltung.

8.1.3 Sicherheitsmanagement

In Bezug auf das ISMS von Hochrisiko Systemen ist als wichtige Beobachtung die Bedeutung des ISMS-Teams und die Verantwortung durch die Leitungsebene hervorzuheben. Im BSI Grundschutzkompendium sind die möglichen Gefährdungen für die Informationssicherheit im Allgemeinen bereits unter den entsprechenden Bausteinen in der Schicht ISMS und Organisation und Personal (kurz ORP) festgehalten (9) (BSI, 2023, ORP). In engeren Fokus auf KI ist hier die Bedeutung der XAI-Expertin im Rahmen des ISMS zu betonen. Wenn diese Person, die beauftragt wird, ein Risikomanagement wie das hier entwickelte zu realisieren, dann muss sie mindestens in den relevanten Teilbereichen der hier veranschlagten Disziplinen geschult werden. Dies umfasst Kenntnisse im Bereich maschinellen Lernen und Datenanalyse, Statistik, als auch insbesondere Ethik, Ontologie komplexer Systeme, Digitalgesetzgebung und Informationssicherheitsmanagement. In der Praxis muss sich diese Person als integre, vertrauenswürdige Ansprechpartnerin für von KI-Systemen betroffenen Personen qualifizieren. Für größere Institutionen, die Hochrisiko-KI-Systeme für ihre Kerngeschäft verwenden, empfiehlt es sich hierzu eigene Stellen zu schaffen. Damit im Zusammenhang sollten in Anlehnung an Baustein ORP.3 umfassende Sensibilisierungs- und Schulungsmaßnahmen für alle Beschäftigten zum Einsatz kommen, die mit Hochrisiko-KI-Systeme interagieren (BSI, 2023, ORP. 3). Das ISMS-Team, der ISB und die XAI-Beauftragte wären in der Verantwortung entsprechende Schulungen und Workshops zu planen, durchzuführen und zu evaluieren. Als zweite Beobachtung ist festzuhalten, dass es sich empfiehlt Gütekriterien dieser Art wenn möglich schon an den bestehenden Standards und Zertifizierungen zu orientieren, damit diese für die Praxis eines ISMS hilfreich werden. Das BSI-Grundschutzkompendium ist zuletzt im Jahre 2023 aktualisiert worden. Eine Weiterentwicklung ist für das Jahr 2026 angekündigt worden.[76] Entsprechend gibt es auch noch keinen eigenen Baustein *Künstliche Intelligenz* im Kompendium. Gütekriterien wie die vorliegenden könnten nun zu einem solchen Baustein modelliert werden. Diese sollten entsprechend bereits die Bestimmungen aus der KI-Verordnung und der DSGVO, sowie weiterer Compliance Anforderungen berücksichtigen. Der Baustein könnte

[76] Dies wurde noch nicht offiziell bestätigt (Reinhardt, 2025).

neben der Informationssicherheitsbeauftragten auch die Schaffung einer XAI-Beauftragten vorsehen, analog zur Schaffung einer ICS-Beauftragten für die erhöhten Anforderungen von Industriekomponenten (BSI, 2008, S. 45f.). Damit würde dem erhöhtem Risiko durch Blackbox Systemen von behördlicher Seite gerecht werden. So wie die ICS-Beauftragte den erhöhten Schutzbedarf in industriellen Steuerungssystemen institutionell absichert, sollte eine XAI-Beauftragte den besonderen Risiken von Blackbox-KI-Systemen begegnen. Die Gütekriterien könnten darüber hinaus auch zur Erstellung von Rechts- und Ethikgutachten genutzt werden, um die Leitungsebene und das ISMS-Team bei der Risikoklassifikation und dem Ergreifen entsprechender Maßnahmen in Bezug auf Interpretierbarkeit zu unterstützen.

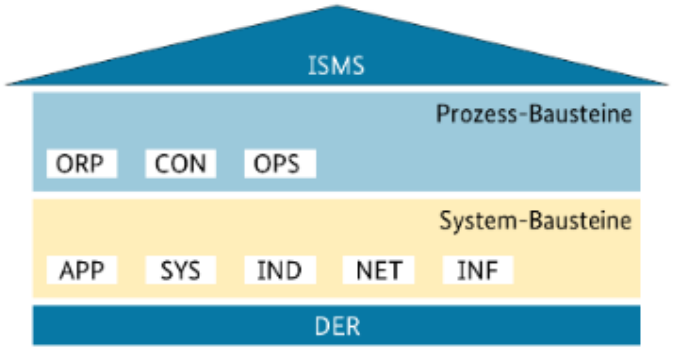

Abbildung 9: Schichtenmodell BSI-Grundschutz

Quelle: Darstellung nach (BSI, 2023, S. 1)

8.2 Diskussion

Die Diskussion umfasst zwei Komponenten, einmal eine engere Diskussion, die sich mit dem angewandten Teil der Arbeit beschäftigt, das heißt vor allem den Gütekriterien und der Evaluation. Diese ist als direkte Erweiterung zu den Gütekriterien zu lesen, da die Abschnitte in konkreten Vorschlägen zu Modifikationen, Erweiterungen und Alternativen münden. Der zweite Teil der Diskussion wagt, den philosophischen Anspruch der Arbeit gemäß, eine vielleicht etwas unorthodoxe Reflexion auf die Voraussetzungen und Grenzen unserer Verstehens- und Vertrauensbemühungen in einer hyperkomplexen Wirklichkeit. Dabei dient die gesamte Diskussion uns *begrifflich* für das ontologische Verhältnis vom *Allgemeinen* und *Konkretem* in Bezug auf diese Gütekriterien zu sensibilisieren.

8.2.1 Skalierung der Komplexität und Emergenz

Unter (6.2) wurde die Herausforderung der wachsenden Modellkomplexität angesprochen. Die Leistungssteigerungen von Deep Learning Ansätzen sind allgemein gesprochen stark mit der Skalierung der Datensätze, der Modellparameter, den computational Ressources und fortgeschrittenen Trainingsmethoden wie RLHF korreliert. Durch das weitere Skalieren dieser Komplexität im Kontext der sogenannten generativen KI-Systeme (vor allem der Large Language Models) wird der Verstehens- und Vertrauensverlust weiter dramatisiert. Dies liegt neben den bereits angesprochenen Aspekten darin begründet, dass es einiges an experimenteller Evidenz gibt, dass durch Skalierung und Training emergente Fähigkeiten („emergent abilities“) auftauchen (Berti, Giorgi & Kasneci, 2025; Wei et al., 2022). Dabei liegt diesen experimentellen Forschungen kein stark einheitliches Verständnis von Emergenz zugrunde. Einige Arbeiten beziehen sich auf die schiere Größe der Modelle und beobachteten, dass bestimmte Fähigkeiten von generativen KI-Modellen bis zu einer kritischen Schwelle nahezu random sind, doch ab dieser Schwelle machen diese einen qualitativen Sprung (zum Beispiel im Bereich Arithmetik oder Chain-of-thought: Math word problems). Dabei ist insofern von emergenten Fähigkeiten zu sprechen, dass diese Performance nicht prognostiziert werden kann, indem die Fähigkeiten kleinerer Modelle extrapoliert werden (Wei et al., 2022). Andere Arbeiten entdeckten bereits bei relativ kleinen Modellen die Emergenz von *impliziten Repräsentationen* der Welt. Zum Beispiel konnte gezeigt werden, dass GPT-Modelle, die nur auf synthetischen Sequenzen legaler Züge eines Brettspiels trainiert wurden und weder in den Trainings- noch Kontextdaten Informationen über die Regeln oder das Spielbrett enthalten waren, legale Züge vorhersagen konnten. Ein MLP-Decoder wurde mit den Hidden-States des Modells trainiert und konnte daraufhin das gesamte Spielfeld, nicht nur den nächsten Zug, rekonstruieren, was auf eine interne Repräsentation des Spielbretts hinweist. Weiterhin wurden die gelernten Hidden States des Modells gezielt manipuliert, woraufhin das Modell nur noch Züge vorhersagte, die mit der Manipulation der gelernten Repräsentation konsistent sind. Hier ist der qualitative Shift darin zu sehen, dass die Modelle nicht, wie es oft angenommen wurde, oberflächliche sogenannte „surface statistics“ lernen, sondern tatsächlich eine *interne Weltrepräsentation* ausbilden, die einen *messbaren kausalen Einfluss* auf die Prognosen und Aktionen der Modelle haben (Li et al., 2024). Damit wäre die oftmals zitierte, scharfe Kritik an den großen Sprachmodellen falsifiziert, die diese als *bloße* stochastische Papageien vorstellt (Bender

et al., 2021). Bevor ich auf die Implikationen zu sprechen komme, ist noch eine wichtige Einschränkung hinzuzufügen. Der Begriff Emergenz verweist hier auf Eigenschaften von generativen KI-Modellen, die wir nicht problemlos antizipieren können. Soweit ich dies in der Analyse der Literatur ermitteln konnte, handelt es sich bei diesen Eigenschaften dennoch höchstens um eine Form der schwachen Emergenz. Die hier evolvierenden Eigenschaften implizieren eine Verstehensgrenze der Klasse 1 (5.7). Das bedeutet, diese Eigenschaften zu prognostizieren übersteigt womöglich unsere technischen und kognitiven Ressourcen. In diesem Sinne sind diese Eigenschaften komplex. Sie sind aber nicht hyperkomplex, wovon wir im Falle einer starken Emergenz sprechen müssten, das heißt evolvierende Eigenschaften *sui generis*. Dies wäre etwa der Fall wenn wir echte willkürliche, interessengeleitete Handlungen bei KI gestützten Agenten erleben würden. In diesem Zusammenhang ist dieser Tage im Zusammenhang mit Generativen KI-Agenten viel von autonomen Agenten die Rede. Obgleich meines Erachtens die Rede von autonomen KI-Agenten etwas irreführend ist, wie sie von vielen Autor:innen betrieben wird. Denn im Rahmen der hier vorgestellten Anthropologie ist es eine notwendige Bedingung für Autonomie (aber keinesfalls eine hinreichende), dass Wesen längerfristig, motiviational-intentional zielorientiert handeln. Von einer motivationalen und intentionalen Zielverfolgung kann jedoch bei keinem der fortgeschrittenen Reasoning Modelle die Rede sein. Die Möglichkeit zur Selbstkorrektur, Feedback-Schleifen und dem Ausführen, selbst komplexer Aktionen im Agentenmodus ist davon immer noch qualitativ unterschieden. KI-Systeme haben keine Motivationen, da sie letztlich dezentral organisierte, datenverarbeitende Maschinen sind, denen wesentlich das *phänomenale Bewusstsein* in einem *lokal verkörperten Organismus* fehlt (siehe auch (Gabriel, 2016c)). Die relevante Implikation ist, dass durch das Anwachsen der Komplexität (Training, Daten, Parameter) starke qualitative Veränderungen wahrscheinlicher werden und aufgrund eben dieser Komplexität lassen sich nur eingeschränkt prognostizieren. Im Gegenteil, die Forschungsgemeinschaft ist immer wieder aufs Neue von den emergenten Fähigkeiten generativer KI-Systeme überrascht. Dafür steht bezeichnenderweise, dass sich ein ganzes Forschungsfeld ausgebildet hat, welches sich überwiegend mit der experimentellen Entdeckung der Fähigkeiten von großen Sprachmodellen in einem quasi-behavioristischen, post-hoc Setting beschäftigt und damit die Modelle weitgehend als Blackboxen behandelt (beispielsweise (Brown et al., 2020; Kojima et al., 2023; Radford et al., 2019)). Das epistemische Sicherheitsrisiko wird durch diese Trends weiter radikalisiert, denn das Potenzial

von nicht antizipierbaren und nicht experimentell zu prüfenden Risiken wächst damit weiter an. Wenn solche Systeme zum Beispiel in relevante Geschäftsprozesse, kritische Infrastrukturen oder der Verarbeitung vertraulicher Informationen integriert sind, dann wäre unter anderem ein Data bzw. Modell-Poisoning Attack, der diese Eigenschaften ausbeutet, denkbar. Bei einem solchen könnten durch den Poisoning-Attack eine interne Repräsentation ausgebildet werden, die bei großen Modelle nur schwerlich in der Testung gefunden wird, deren Funktionsweise später als Backdoor für die Angreifer wirksam wird, wodurch sie sich Zugang zu kritischen Informationen und Prozessen beschaffen könnten. Vor dem Hintergrund dieser unkontrollierbaren Komplexität ist insbesondere die Integration von großen Sprachmodellen in immer mehr Geschäftsabläufe durchaus bedenklich (für verwandte Angriffsvektoren siehe (BSI, 2024b, 2025)).

Vorschlag: Die Gütekriterien sollten in kommenden Projekten von der *spezifischen KI* auf die *systemischen Risiken* durch die sogenannte *General Purpose AI* wie Transformer Modelle erweitert werden. Ein erster Ansatz wäre mittels einer intensiven Literaturrecherche eine Taxonomie zu entwickeln, welche die bestehenden emergenten Fähigkeiten mit ihren Risiken gegenüberstellt, um diese in das Risikomanagement zu integrieren. Offizielle Stellen wie das EU-AI-Office, das BSI oder die Bundesnetzagentur könnten entsprechende Taxonomien nach dem Vorbild der Grundschutz-Bausteine entwickeln und veröffentlichen. Dabei könnten gezielte Tests auf diese bekannten Gefährdungen durchgeführt werden und den Beteiligten die wachsende Gefahr durch Komplexität vermittelt werden.

8.2.2 Fall-spezifische Evaluation und Validierung

Die Lektüre dieser Arbeit legt nahe, dass für ihr Verfassen tendenziell Risikoszenarien wie die Kreditrisikoanalyse oder verwandte Risikoszenario als Modell gedient haben. Bei diesen geht es erstens darum, dass ein Subjekt S_x auf der einen Seite von dem Output einer Blackbox KI auf der anderen Seite betroffen ist und die erklärungsgebende Instanz V eine autonomiefördernde Vermittlungsrolle einnimmt. Und zweitens geht es immer darum, dass durch den Modelloutput über die Zuweisung von Ressourcen entschieden wird, also Kredite, Hochschulzugänge, medizinische Versorgung, Versicherungen und dergleichen. Aus diesen konkreten Fällen wurden dann relativ allgemeine Gütekriterien gewonnen. Bei der weiteren Verwendung dieser Kriterien sollten wir nun eine Art von induktivem Fehlschluss vermeiden. Auch wenn wir in einer Serie von Evaluationsstudien eine Vielzahl

von konkreten Modellen evaluieren und mit den gewonnenen Daten die Gütekriterien und den Ablaufplan weiter modifizieren, bleibt es dennoch unrealistisch davon auszugehen, dass alle relevanten Einzelfälle evaluiert werden. Im Gegenteil, die Praxis der Anwendung von KI-Modellen übersteigt immer die Modellierung. Konkret bedeutet dies zum Beispiel, dass die Gütekriterien und Metriken in jedem Einzelfall anders angelegt werden müssen. Zum Beispiel könnte der Schwellenwert für Suffizienz im Falle eines Triage Modells deutlich höher liegen als bei der Kreditrisikoanalyse. Aus diesem Sachverhalt lässt sich ein weiteres wichtiges Desideratum ableiten (Mothilal & Tan, 2021).

Vorschlag: Jeder neue Fall S und jedes Modell M stellen wichtige Datenquellen für die Entwicklung von sicheren und interpretierbaren KI-Systemen dar. Die konkreten Fälle sollten den allgemeinen Ablaufplan, die Methoden, Metriken und so weiter modifizieren und vice versa.

8.2.3 Gütekriterien als Ethik der Differenz

Das Verhältnis vom Allgemeinen zum Konkreten wird nochmal besonders relevant, wenn wir die Diskussion nun auf die Betroffenen richten. Für diese Gütekriterien wurde ein abstrakter Begriff von Autonomie und informationeller Selbstbestimmung als schützenswertes Gut gewonnen, welches auf ein nicht minder abstraktes epistemisches Subjekt S_x hin zentriert worden ist, um letztlich in einem steril, bürokratisch anmutenden Ablaufplan zu münden. Dieses Vorgehen ist ein durchaus sinnvoller Ausgangspunkt, doch sollten die Gütekriterien hier nicht stehen bleiben, wenn sie zu adressieren beanspruchen, was für menschliche Autonomie wirklich von Bedeutung ist. Denn was für Menschen von Bedeutung ist, kann nicht hinreichend aus dem *allgemeinen* philosophischen Begriff, sondern muss aus den *konkreten* lebensweltlichen Vollzügen eines jeden Menschen heraus verstanden werden. Daher gilt es in diesem Teil diese Abstraktionsleistung ein Stück weit zu revidieren, um das abstrakte Subjekt S_x wieder lebensweltlich zu verorten.

Habitus und Diskriminierung In dieser Lebenswelt sind Entscheidungen über ökonomische, soziale, politische und sonstige Teilhabe zunehmend durch datenverarbeitende Systeme mitbestimmt. Dies führt schon seit langem zur systematischen Reproduktion und Amplifizierung von Diskriminierungsmustern, zu denen unter anderem Sexismus, Rassismus und Klassizismus, sowie insbesondere auch intersektionale Diskriminierung zu zählen ist.

Menschen werden hier aufgrund ihres Geschlecht, ihrer Hautfarbe, ihrer sozialer Herkunft, ihres Alter, ihrer religiösen oder politische Weltanschauung diskriminiert (Caspar, 2023; D'Ignazio & Klein, 2023; Friedman & Nissenbaum, 1996; Mühlhoff, 2023a; S. U. Noble, 2018). Diese öffentliche Komponente dieser Marker bildet bei jedem Individuum (als auch Gruppen, Milieus und Schichten) eine historisch kontingente Konstellation von Merkmalen aus, kurz der *Habitus* (Bourdieu, 1984; Fröhlich & Rehbein, 2014, S. 110ff.). Der Habitus ist ein Arrangement solcher Merkmale und bestimmt über Stellung und Status in der Gesellschaft. Die epistemische Dimension der Diskriminierung, etwa des Rassismus, begeht nun den Fehlschluss den Habitus als soziale Identität zu einer ontologischen Identität zu essenzialisieren (Gabriel, 2021, S. 197ff.). Da hierdurch problematische Diskriminierungsmuster reproduziert werden, lädt diese Beobachtung vielleicht zunächst dazu ein, zugunsten der Gleichbehandlung aller Betroffenen von all diesen Merkmalen zugunsten von S_x zu abstrahieren und wie in der Evaluation vorgenommen ausschließlich ökonomische Merkmale im engeren Sinne zu betrachten. Hier sind nun mehrere Probleme zu nennen. Erstens ist hier das Problem der Proxy-Diskriminierung zu betonen (Stanford Encyclopedia of Philosophy, 2025; Tschantz, M. C., 2022). Gerade in Gesellschaften mit einer sehr stark ausgeprägten sozioökonomischen Ungleichheit in Bezug auf Bildung, Einkommen und Vermögen wie der Bundesrepublik ist eine solche Proxy-Diskriminierung, zum Beispiel der Herkunft oder des Geschlechts durch Variablen wie höchster Bildungsabschluss oder Jahresbruttoeinkommen, oft wahrscheinlich.[77] Zweitens abstrahiert die vermeintliche Gleichbehandlung des Subjekts S_x von historisch gewachsenen Ungleichheiten und Ungerechtigkeiten, wie dem dem Gender Pay Gap oder Benachteiligung aufgrund der Hautfarbe.

Gedankenexperiment HR KI-Assistent In einem einfachen Gedankenexperiment können wir uns statistisch vor Augen führen, welche moralischen Konsequenzen eine solche Abstraktion mit sich bringt. In diesem Gedankenexperiment sind alle Führungspositionen deutscher Unternehmen seit Generationen ausschließlich von Männern besetzt. Nun ist eine Stelle für eine Führungsposition vakant. Auf diese Stelle bewerben sich sowohl Frauen als auch Männer. Für das Bewerbungsverfahren wird ein Human Ressource KI Assistent

[77] Für die Bildungsungleichheit sind unter anderem die großen Bildungsstudien in den Blick zu nehmen (IGLU/PIRLS Germany, 2025; Institut zur Qualitätsentwicklung im Bildungswesen (IQB), 2022; Lewalter et al., 2023; Organisation for Economic Co-operation and Development (OECD), 2023). Für Einkommens- und Vermögensbezogene Ungleichheit sind die Studien der renommierten wirtschaftswissenschaftlichen Institute wie das DIW zu empfehlen (DIW Berlin, 2020), aber auch (Linartas, 2025) und natürlich (Piketty, 2014) und (Riddell et al., 2024).

eingesetzt. Das System wird aufwendig dahingehend entwickelt, dass es soziodemographische Merkmale wie Alter, Herkunft oder Geschlecht ignoriert. Aufgrund der historisch gewachsenen Verhältnisse ist von folgender Ausgangslage auszugehen:

- Es bewerben sich deutlich mehr Männer als Frauen auf die Stelle.
- Die Frauen sind sozioökonomisch schlechter gestellt als die männlichen Mitwerber (das heißt niedrigeres Bildungsniveau, weniger gefragte Abschlüsse usw.).
- Die weiblichen Bewerbenden haben weniger Berufserfahrung.
- Das Modell abstrahiert von diesen biographischen, geschlechtlichen und sonstigen demographischen Merkmalen.
- Das Modell ist überwiegend (oder ausschließlich) mit männlichen Erwerbsprofilen trainiert worden.

Das Modell wurde nun nicht explizit auf der Variable Geschlecht trainiert. Doch da überwiegend (oder ausschließlich) männliche Profile in dem Trainingsdatensatz enthalten sind, prägt es einen solchen Proxy-Bias aus. Hinzu kommt der Aspekt, dass sozioökonomische Variablen fokussiert werden, die tendenziell mit männlichen Erwerbsbiographien assoziiert sind. Dieses Gedankenexperiment ist zugegeben sehr stilisiert, doch veranschaulicht es die relevante Schlussfolgerung. Erstens ist es in dieser Ausgangslage prognostisch unwahrscheinlich, dass es eine weiblich gelesene Person eine Stelle für eine Führungsposition bekommt. Das bedeutet, dass trotz der scheinbaren Gleichbehandlung die historisch gewachsenen Ungleichheiten (und ggf. Ungerechtigkeiten) in diesem Szenario wahrscheinlich nicht abgebaut werden können. Erweitern wir diese Grundüberlegung auf die gesamte Gesellschaft führt eine solch abstrakte Betrachtung von Sicherheit, Autonomie und Selbstbestimmung zur Perpetuierung und Verkrustung gesellschaftlicher Verhältnisse. Zweitens kommt auch hier wieder eine normative Komponente ins Spiel. Als Teil des Feature Engineerings und des Labelns der Zielkategorien wird entschieden, ob diese Features überhaupt von dem Modell als relevant gelernt werden können. Es ist eine Entscheidung, dass gerade diese, historisch mit männlichen Erwerbsbiographien assoziierten sozioökonomischen Variablen eine entsprechend hohe Gewichtung bekommen. Die daraus resultierenden Selektionen folgen nicht aus der Natur dieser Unternehmen oder Struktur der Marktwirtschaft, sondern

sind eine normative Entscheidung von Entwickelnden und anderen Involvierten.

Identität und Differenz An dieser Stelle können wir zwei Ausprägungen von Methoden unterscheiden, wie auf diesen Missstand reagiert werden kann. Der ersten Ausprägung können wir den Titel Identitätspolitik und der Zweiten den Titel Differenzpolitik geben. Die Identitätspolitik besteht in ihrer stärksten Ausprägung darin, soziale Merkmale zu einer Identität zu essenzialisieren und daraus bestimmte politische Forderungen abzuleiten. Paradigmatisch stehen hierfür die Verschmelzung von sozialen Merkmalen zu Identitäten von vermeintlich benachteiligten Gruppen, wie der *alte weiße Mann*, *die Frauen* oder *die Migrant:innen* (Gabriel, 2021; Sen, 2006, S. 249ff.). Doch die Identitätspolitik fußt auf fragwürdigen Prämissen (Gabriel, 2021, S. 244ff.). Denn Menschen werden bestimmte Ressourcen und Chancen zu- oder abgewiesen aufgrund ihre Zugehörigkeit zu bestimmten Gruppen. Die sachlich korrekte Beobachtung der Identitätspolitik ist die Unterschiedlichkeit des Menschen. Doch die Differenz zwischen Menschen geht tiefer. Dies können wir uns nochmal am Beispiel der Erwerbsbiographien veranschaulichen. Es gibt bestimmte soziodemographische und -ökonomische Merkmale, die stärker mit männlichen oder weiblichen Personen korreliert sind. Doch jede Biographie, sowie die Identität eines Menschen ist eine individuelle Konstellation von psychologischen, sozialen, biographischen, biologischen und sonstigen Eigenschaften. Es gibt nicht *die* weibliche Erwerbsbiographie, sondern nur eine statistische Häufung von Korrelationen, die wir dann in unseren alltäglichen, aber auch wissenschaftlichen Urteilen *ad hoc* zu bestimmten Identitäten essenzialisieren (Lippmann, 2018, S. 127). Wenn *die* Frauen beispielsweise durch einen Geschlechtsbias bevorzugt würden, würde wieder von der individuellen Lebenslage jeder einzelnen Person, jeder Frau in ihrem individuellen Menschsein abstrahiert werden. Doch weder gibt es *die* Frauen, noch *den* alten weißen Mann im eminenten Singular, als eine ontologische Kategorie, genau sowenig wie es *die* Afroamerikaner gibt. Machen wir diese Überlegung nochmal am konkreten Fall einer Person deutlich, die sich durch diverse Hürden nicht selbstständig aus der Armut befreien kann. Dabei handelt es sich um eine von Armut betroffene, alleinerziehende, weiblich gelesene Person ohne Berufsabschluss, die im sogenannten Niedriglohnsektor arbeitet. Außerdem muss sie, wie in Deutschland nicht ungewöhnlich, pflegebedürftige Angehörige pflegen. Aufgrund der mangelhaft ausgebauten Betreuungsinfrastruktur für Kinder (Kitas und Ganztagsschulen) und dem Mangel an formaler Pflege und der körperlichen, psychischen und wirtschaftlichen Belastung ist die Person nicht in der Lage ihr Humankapital

auszubauen und beispielsweise eine Ausbildung oder eine Abendschule zu besuchen (DIW Berlin, 2014, 2024). Diese individuelle Biographie ist gar nicht sinnvoll erfasst mit der statistisch zu grobkörnigen Kategorie des Geschlechts. Sozioökonomisch teilt diese Person mutmaßlich deutlich mehr Merkmale mit einigen männlichen Hauptschulabsolventen in Deutschland als mit vielen weiblich gelesenen Personen. Bei der Mehrheit dieser Hürden ist es realistisch davon auszugehen, dass es sich um strukturelle, von der Person unverschuldete Bedingungen handelt, die sie daran hindern, sich aus Eigeninitiative aus der Armut zu befreien. Aus der Perspektive der Differenzethik stellt sich dann die Frage, wie alle gesellschaftlichen Akteure die von dieser Person unverschuldeten Hürden ausgleichen können (Gabriel, 2021, S. 256f.). Konsequent zu Ende gedacht, bedeutet dies für den Arbeitsmarkt und Bewerbungsverfahren, dass eine Art positive Diskriminierung aller Frauen gar nicht allen Frauen in ihrer individuellen Lebenslage gerecht würde. Stattdessen müssten individuelle Verfahren und Programme die soziale Benachteiligung dieser Person ausgleichen. Universalismus bedeutet somit nicht, dass alle Menschen gleich sind, was manchmal auch abwertend unter dem Slogan "Gleichmacherei"kursiert. Stattdessen bedeutet Universalismus, dass alle Menschen gleich darin sind, dass sie anders sind, sie sind gleich in ihrer Individualität, was in dem lateinischen Ausdruck *Individuum* für *Unteilbares* anklingt (Gabriel, 2021, S. 249ff.). Universalismus bedeutet, dass alle Menschen es *gleichermaßen* verdienen als Individuum in ihrer sozialen, demografischen, biographischen, historischen, leiblichen und geschlechtlichen Position behandelt zu werden und dementsprechend nicht als stereotypische Repräsentanten oder Instanzen von Identitäten (siehe auch (Boehm, 2022)).

Implikationen für Hochrisiko Systeme Und dies müssen wir auch immer beim Entwickeln und Evaluieren von entsprechenden Gütekriterien berücksichtigen. Aus diesem Grund kann das abstrakte epistemische Subjekt S_x vielleicht das erste Wort bekommen, aber nicht das letzte. Andernfalls tun wir dem Menschen begriffliche Gewalt an.[78] Es würde den Rahmen dieser Arbeit sprengen, doch es wäre ein lohnenswertes Projekt den philosophischen Begriff einer Differenzethik technisch und formal auf die Entwicklung von KI-Modellen und dem Anwenden dieser Gütekriterien hin zu übersetzen. Ein Weg wäre eine Taxonomie dieser Gütekriterien für eine Vielzahl von Fallkontexten zu entwickeln,

[78] Siehe zu diesem Themenkomplex auch die Arbeit über den Zusammenhang von Identität und Gewalt (Sen, 2006).

um Anwendende Leitfäden für diverse Fälle an die Hand zu geben.

Empfehlung: Mittel- und langfristig empfiehlt es sich moralisch jeden Menschen als Menschen und nicht als typischen Repräsentanten einer Gruppe bzw. einer Identität zu behandeln. Eine der leitenden Fragen könnte sein: *Was schulden wir dieser Person in ihrem individuellen Lebenskontext?* Und daraus folgt, dass jedes Profil individuell bearbeitet werden sollte. Daraus folgen noch zu entwickelnde Anforderungen für die Entwicklung von Hochrisiko-KI-Systemen und insbesondere die Verantwortung im Forum jeden Fall in einen biographischen Kontext zu verorten.

8.2.4 Optionale Methoden

Diese Arbeit ist analytisch und formal vorgegangen und hat letztlich argumentativ einen Interpretierbarkeitsbegriff entwickelt. Die normative Grundbestimmung des Menschen als freiheitliches-autonomes Wesen ist eine philosophische Prämisse, die juristisch validiert werden kann. Wenn wir diese Normativität ins Zentrum unserer Analyse rücken, dann leiten sich daraus entsprechende, strenge Anforderungen an vertrauenswürdige KI-Systeme ab. Ein anderer Ansatz wäre experimenteller Natur und würde den normativen Ausgangspunkt dadurch gewinnen, dass Menschen unter verschiedenen Szenarien befragt werden, welche Anforderungen sie an Erklärbarkeit stellen (Miller, 2019). Die so ermittelten Gütekriterien könnten dann verglichen werden mit dem Rechtsverständnis und dem Menschenbild, welches sich aus den hier zur Grundlage genommen Texten ergibt. Dennoch ist damit die Option echter sozialer Entfremdung nicht vom Tisch. Das heißt, wenn Menschen nach bestimmten statistischen Mittelwerten *empirisch* bestimmte Anforderungen an Erklärungen stellen, folgt daraus nicht notwendigerweise, dass dies die Anforderungen sind, die wir auch stellen *sollten*. Dies wäre eine Variante eines naturalistischen Fehlschlusses, der vom Sein aufs Sollen schließt (Ridge, 2025). Beide Ansätze haben ihre Vorteile und sollten meines Erachtens noch stärker in der Forschung komplementär gedacht werden. Jedoch sollten wir uns aus philosophischen Gründen letztlich immer davor hüten, von der Deskription wie Menschen sich de facto verhalten, präskriptiv darauf schließen, was wir tun *sollen*.

8.2.5 Prädikative Analytik und informationelle Integrität

Das KI-Zeitalter ist immer noch primär von einem anthropologischen Paradigma bestimmt, welches wir in Anlehnung an Meyers Formulierung knapp auf die Formel „Subjekt als

System“ bringen können (Meyer, 2024, S. 36). Ein Wesen als System zu konzeptualisieren, bedeutet, verkürzt gesagt, dieses ontologisch als etwas zu verstehen, was sich vollständig in den Begriffen der empirischen Wissenschaften beschreiben lässt. Und dies bedeutet insbesondere, dass sich das Verhalten von Systemen mittels objektiver Gesetze erklären und vorhersagen lässt. Dies meint im Kontext von prädikativen Analysen durch (Hoch-)risiko-KI-Systeme, dass der Mensch, wenn auch methodisch mitunter anspruchsvoll, letztlich als ein System betrachtet wird, für dass sich empirische Gesetzmäßigkeiten entdecken lassen, welche den Menschen *hinreichend* beschreiben. In dieser Optik wird der Mensch letztlich doch auf ein Bündel statistischer und logischer Gesetzmäßigkeiten reduziert, welche sich mit Hilfe der positiven Wissenschaften (psychologisch, soziologisch, ökologisch und so weiter) entschlüsseln lassen. Doch dieser Anthropologie widersprechen zwei zentrale Argumentationslinien, die oben schon angerissen wurden (4).

1. Erstens implizierte die diskursive Vernunftnatur das Vorhandensein echter offener Möglichkeiten für menschliches Handeln. Die Grundstrategie würde ich vorsichtig und in aller Kürze wie folgt rekonstruieren. Menschen müssen sich notwendigerweise als solche Wesen verstehen, die als Teilnehmende in vernünftigen Diskursen um die besseren Argumente ringen. Und dies wiederum impliziert die Existenz echter, offener Möglichkeiten für unser Urteilen und Handeln (Meyer, 2024, S. 30). „Ein Subjekt ist nicht nur ein System, in dem etwas passiert, sondern es ist auch jemand, der etwas entscheidet und tut, und sein Tun ist nicht nur eine Art und Weise, wie es ein bestimmtes Geschehen erlebt.“ (Meyer, 2024, S. 33) Dies entspricht auch der Praxis der empirischen Wissenschaften, wie etwa der empirischen Psychologie (Meyer, 2024, S. 57). Auch die hier mobilisierten Disziplinen wie Informationssicherheit und Recht und die entwickelten Gütekriterien sind in ihrer Praxis, wie die Ausführungen zeigten, auf diese diskursive Struktur der Vernunft angewiesen.

2. Das Soziale ist das Ergebnis einer Integration von letztlich irreduzibel differenten Perspektiven individueller Lebewesen (Dissensmanagement). Menschen sind in diesem Sinne konstitutiv sozial produzierte Lebewesen (Gabriel, 2020a, S. 427ff.). Dabei bleibt letztlich eine Unterhintergehbarkeitsthese, dass bedeutet, dass der (sozialisierte) Standpunkt eines jeden Menschen irreduzibel der *Jemeinige* ist, um an ein geflügeltes Wort Heideggers zu erinnern (Gabriel, 2020a, S. 268ff., S. 461ff.). Dabei verwende ich

> den Ausdruck *mutatis mutandis* als Beschreibung für die letztlich nicht positivierbare Individualität eines jeden Menschen. Ich würde dies in diesem Zusammenhang so deuten, dass die sozialisierte Perspektive zweier Menschen, das bedeutet ihr indexikalisches Erleben der Wirklichkeit, ist niemals identisch, was überhaupt erst die Voraussetzung der Sozialität ist. Jemeinigkeit und Sozialität sind keine Widersprüche, sondern notwendig komplementär. „Zum existierenden Dasein gehört die Jemeinigkeit als Bedingung der Möglichkeit von Eigentlichkeit und Uneigentlichkeit. Dasein existiert je in einem dieser Modi, bzw. in der modalen Indifferenz.“ (Heidegger, 1927/1977, S. 41–42)

In diesem Sinne tut ein statistischer Positivismus, der das Zeitalter datenverarbeitender Systeme prägt, dem Menschen in seiner individuellen Dignität immer Gewalt an.[79] Und das ganz gleich wie feinkörnig die Features und Profile eines Kreditrisikosystems aufgeschlüsselt werden. Das liegt dann letztlich darin begründet, dass es den Menschen als ein solches prognostizierbares System gar nicht gibt bzw. nur in dem Sinne gibt, dass es sich dabei um eine durchaus wirkmächtige *Konstruktion* handelt, eine falsche Repräsentation der menschlichen Lebensform als Element eines Paradigmas, welches im Hintergrund des Maschinellen Lernens und der KI-Forschung aktiv ist. Aber durch dieses Paradigma wird dem Menschen *per Design* die Möglichkeit echter offener Möglichkeiten des Handelns sowie sein unhintergehbarer individueller Standpunkt beraubt. Mit diesen beiden Argumenten bekommt die jüngere rechtswissenschaftliche und rechtsethische Debatte weiteren Rückenwind, die über ein Recht auf informationelle Selbstbestimmung hinaus, ein Recht auf informationelle Integrität fordert. In diesem Kontext ist das Recht auf informationelle Integrität so zu interpretieren, dass Menschen ein Recht darauf haben sollten, *nicht als System* betrachtet zu werden. Stattdessen sollten Menschen als Wesen betrachtet werden, die individuell differente praktische Identitäten ausprägen und in Kommunikation mit diesen die echte Möglichkeit haben, den Lauf ihrer individuellen Biographie zu verändern, das heißt zu handeln.

Empfehlung: Als Gesellschaft täten wir gut daran, Institutionen und Technologien zu bauen und zu fördern, die dieser Norm der informationellen Integrität gerecht werden. Das bedeutet einmal mehr, das Forum nicht als Appendix einer zuvor stattgefundenen

[79] In einem ganz ähnlichem Zusammenhang spricht Paul Schütze treffend von „Mining the Future“ (Schütze, 2022).

Datenauswertung für eine Kreditprognose zu betrachten, sondern als integraler Bestandteil der Prognose. Und es bedeutet, dass das metaphysische Paradigma Subjekt als System nicht konstitutiv werden sollte, für Gesetzgebung und Rechtsprechung. Im Gegenteil, das Forum sollte einen echten Diskurs realisieren, das heißt einen Ort, der den Verlauf der Dinge (Kreditentscheidung, Bewerbungsverfahren) verändern kann. Die DSGVO geht bereits Schritte in die Richtung informationeller Integrität, so ist es in Anlehnung an den Artikel 22 davon abzusehen, über die Zuweisung gesellschaftlicher Güter, wie zum Beispiel Arbeitslosengeld oder Ausbildungsförderungen wie das bundesdeutsche Bafög *alleine* auf der Basis datenverarbeitender Systeme zu entscheiden.

8.2.6 Grenzen des individualistischen Paradigmas

Letztlich sind Studien wie die vorliegende sehr Individuums-zentriert. Dies liegt an der Orientierung an lokaler Erklärbarkeit, dem Einzelfall und einem einzelnen zu identifizierenden Subjekt. Dies hat meines Erachtens vor allem rechtliche, philosophische und technische Gründe, da die formalen und quantitativen Methoden solcher Studien intrinsisch einzelfallbasiert sind. Dies ist nicht einfach eine marginale methodologische Entscheidung, sondern ein weltanschauliches Paradigma, welches Verantwortung, Freiheit und Sicherheit individualistisch denkt und (re-)produziert (Mühlhoff, 2023c). Dieses Paradigma hat eine schwere Hypothek. Geht es doch davon aus, dass die Verteilung von Ressourcen und Chancen letztlich, wenn auch unter bestimmten Einschränkungen, basierend auf den (persönlichen) Eigenschaften von Individuen, statistisch ermittelt und entschieden werden *darf*. Wobei die Frage dann wiederum vollkommen unangetastet bleibt, ob es Güter gibt, die nicht Gegenstand eines (digitalisierten), (teil-)automatisierten Selektions- und Zuweisungsmechanismus werden *sollten*. Ein umfassendes Verständnis von informationeller und sozialer Integrität könnte dem entgegensetzen, ob es bestimmte Güter gibt, die Menschen vollkommen unabhängig von ihren persönlichen Eigenschaften (Einkommen, Bildung, Alter und so weiter) zustehen und deren Verwaltung und Zuweisung eine gesamtgesellschaftliche Verantwortung darstellt, das heißt eben eine *kollektive* und keine *individuelle* (Deutschlandfunk, 2022; Mühlhoff, 2023c). Dies scheint eine normative Ausgangslage zu sein, die in unseren Zeiten stärker denn je unter Druck steht, wo selbst das grundgesetzlich zugesicherte Existenzminimum zur Disposition gestellt wird. Wobei diese moralische Innovation, dass Menschen *als* Menschen eine Würde haben, die eben nicht erworben oder verloren werden

kann, vor dem Hintergrund der dunkelsten Zeiten der deutschen Geschichte entstanden ist. Menschen können durch das erwerben oder verlieren von gesellschaftlichen Gütern (Abschlüssen, Titeln, Vermögen usw.) die Menschenwürde selbst nicht erwerben oder verlieren. Die Menschenwürde hält die Idee hoch, dass das Leben als Leben und nichts mehr, selber heilig und schützenswert ist. Entweder haben wir das als Gesellschaft nie gewusst oder wir vergessen es gerade.

8.2.7 Grenzen der instrumentellen Logik

Nach meinem Dafürhalten spricht einiges dafür, dass der moderne Rechtsstaat in gewissem Sinne auf ein instrumentelles Technikverständnis verpflichtet ist. Die Annahme, dass die richtige Ontologie der Technologie eine ist, die diese als Instrumente vorstellt, wurde von der Technikphilosophie und den Sozialwissenschaften zurecht als unzureichend und letztlich auch naiv identifiziert. Es gibt eine Reihe von Argumenten, die diese Ontologie falsifizieren, von denen ich einige wenige meinem Verständnis nach reproduzieren möchte:

1. Es liegt im Wesen von Technologien, ab einer bestimmten Komplexität, und dies gilt für ANNs wie hier gezeigt wurde mit Gewissheit, dass wir vor Inbetriebnahme gar nicht *a priori* antizipieren können, wozu sie fähig sind. Folglich ist eine Beschreibung von Technologie als ein Set von Instrumenten mit endlichen vielen Outputs inkohärent, da wir die möglichen Input-Output Relationen eben nicht überblicken können.

2. Interessanterweise greift die instrumentelle Ontologie auch nicht unter der idealisierten Annahme einer vollständig deterministischen Input-Output-Relation. Dies liegt darin, dass Technologien, die institutionell installiert sind, die Möglichkeit eines zukünftigen Missbrauchs eröffnen. Dieses Argument, welches auch unter dem Namen *Function Creep* diskutiert wird, wird auch gegen Maßnahmen, wie Vorratsdatenspeicherung oder Chatkontrollen vorgebracht. Denn selbst unter der optimistischen Annahme, dass die bestehenden Regierungsverhältnisse und Behörden ihren Datenschatz nicht repressiv ausbeuten, gibt es hierfür keine Garantie in Bezug auf zukünftige Entwicklungen, Regierungswechsel, Gesetzesänderungen und vieles mehr (Leese & Ugolini, 2024; Naarttijärvi, 2022; Pereira & Raetzsch, 2022).

3. Darüber hinaus führen besonders invasive Technologien, wie zum Beispiel Anwendungen von KI oder Messenger zu sozialen Strukturierungseffekten, welche weit über

die instrumentellen Zwecke eines Nutzenden hinausgehen. Ein einfaches Beispiel ist Metas WhatsApp, welches in Deutschland ein beinahe Monopol bei Menschen unter 30 hat (G. Kaiser, 2024; Kim, 2023). Wenn junge Menschen sozial teilhaben möchten, sind sie nahezu auf diesen Anbieter angewiesen. Ein Effekt, welcher als *Lock-in-Effekt* bekannt ist (Farrell & Klemperer, 2007). Wenn junge Menschen dem instrumentellen Zweck, den Messenger für die *Kommunikation* mit Freunden zu nutzen zugestimmt haben, folgt daraus nicht, dass diese auch dem übergeordneten *sozialen Selektionsmechanismus* zugestimmt haben, der sie von der Teilhabe fernhält, wenn sie sich einem Anbieter und seinen teilweise semi-legalen und auch explizit illegalen Datengeschäften entziehen wollen (Caspar, 2023, S. 109ff.), (Haugen, 2023; Mühlhoff, 2018a; Netzpolitik.org, 2024b).

4. Dieses Argument lässt sich zu einer umfassenden sozialontologischen Konklusion erweitern. Technologien spiegeln nicht einfach unsere, sozusagen vor-technologischen Zwecke, sondern sie strukturieren unser soziales Zusammenleben neu und erzeugen damit neue soziale Praktiken, Zwecke, Gewohnheiten, die es *vor* der Einführung und Verbreitung *nicht gab*. Dies lässt sich detailliert anhand der Digitalisierung studieren, die ganz neue Umgangsformen und Verhaltensmuster, zum Beispiel das *Ghosting* oder die mittlerweile nach ICD anerkannte *Gaming Disorder* erzeugt haben. Auch über eine Smartphone-Sucht wird bereits lange diskutiert (Lin et al., 2016; Navarro et al., 2020; Rumpf et al., 2018). Die entscheidende Erkenntnis ist meines Erachtens, dass die menschliche Psychologie keine Menge von Eigenschaften ist, die digital abgebildet wird, sondern digitale Technologien wirken modulierend, selektiv und produktiv auf die menschliche Psychologie.[80]

5. Darüber hinaus sind Technologien nicht wertneutrale Werkzeuge, sondern in ihrer Programmierung, ihrer Haptik, ihrem UX-Design ist auch ein bestimmtes Menschenbild, eine bestimmte Anthropologie eingeschrieben (Gabriel, 2020a; Pörksen, 2000). Paradigmatisch führt das iPhone und dessen nahezu sakrale Aufwertung durch Werbung vor, dass ein vernetztes, leichtes, smartes, datafiziertes Leben ein gutes Leben sei. Es entspräche unserem Wesen, dass das gläserne Selbst durch prädikative Analysen, Nudging und versiegelte Oberflächen, freundlich, aber bestimmend durchs

[80] Für eine umfassende Analyse der Modellierung dieser Mechanismen und ihre Steigerung als „immersive Macht“ im Umfeld von Google siehe die Fallstudie in (Mühlhoff, 2018b).

Leben geleitet wird (Mühlhoff, 2018a).

Trotz all dieser Einwände muss ich gestehen, dass mir unter der Bedingung, dass diese Technologien nun mal zur Anwendung kommen, nicht einfällt, wie wir (natürlich unter gegebenen Einschränkungen und Problematisierungen) vollständig auf eine instrumentelle Ontologie verzichten könnten. Vor diesem Problemhintergrund kann man eine Komplexitätsstudie wie die vorliegende auch so lesen, dass sie diese Ontologie flexibilisiert, diese aber nicht vollständig aufgibt. Es ist meines Erachtens einfach nicht zu erkennen, wie wir unter der Annahme, dass Autonomie eine relevante Größe bleiben soll, vollständig darauf verzichten könnten. Wie ich versucht habe in dieser Arbeit dazulegen, müssen wir stattdessen schonungslos anerkennen, an welcher Stelle Technologien unser normatives Selbstverständnis verletzen, um dann als Antwort auf ihre Komplexität wiederum komplexe Verfahren zu entwickeln, und das heißt eben infrastrukturelle, juristische und behördliche Verfahren, die ihr Risiko minimieren. Um es etwas dramatisch zu sagen, der Geist ist aus der Flasche, jetzt obliegt es der gesellschaftlichen Verhandlung, welches Menschenbild wir durch diese Technologien abbilden, erzeugen und verstärken wollen, eines der Fremd- oder der Selbstbestimmung (siehe auch (Gabriel, 2020a; Mühlhoff, 2023c).

8.2.8 Grenzen der formallogischen Methode

Die Basis dieser Arbeit war die Bedingungsontologie und die sie modellierende formale Logik und Stochastik. Die hier vorgestellte Methode kann dazu verleiten und einige Autoren suggerieren dies sogar recht eindeutig, dass der Ansatz Ereignisse wie die Prognose eines Hochrisiko-KI-Modells formallogisch mit einer Menge von notwendigen und zusammengenommen hinreichenden Bedingungen zu erklären, in einer Hierarchie von formalistischen und normalsprachlichen Ansätzen an der Spitze steht. Das Primat formallogischer Ansätze steht inhaltlich in einem Kontinuum mit dem berühmten Programm zur Entwicklung einer Idealsprache, die vor allen mit den Namen Russell, Frege und Carnap in Verbindung steht (Gabriel, 2016a, S. 120f.). Etwas knapp gefasst, lässt sich dieses Programm so beschreiben, dass durch die logische Analyse der natürlichen Sprachen eine formale Idealsprache gewonnen wird und damit die *ahistorischen Gesetze des Wahrseins* entdeckt (Gabriel, 2016a, S. 121).[81] Dadurch soll insbesondere die vermeintliche Ambiguitäten und Unschärfe der

[81] Dabei handelt es sich bei dieser kurzen Wiedergabe um eine Stilisierung. Die genannten Denker haben zwar an diesem Projekt gearbeitet, doch im Einzelnen und auch diachron durchaus nuancierte Analysen vorgelegt.

normalen oder natürlichen Sprachen (Englisch, Deutsch, Hindi usw.) überwunden werden. Im Ergebnis könnten durch eine Orientierung an die Gesetze dieser Idealsprache Missverständnisse und Konflikte überwunden werden. Die Nuancen und der Facettenreichtum natürlicher Sprachen wird tendenziell in diesem Programm als ein Defizit dieser Sprachen verstanden. In unserem Zusammenhang bieten sich zwei Interpretationen an und beide sind historisch verfolgt worden. Entweder handelt es sich bei diesem hier vorgestellten formalen Ansatz um eine methodische Reduktion mit dem Ziel bestimmte Aspekte auszuklammern, um die Effizienz einer bestimmten Methode zu erhöhen. Oder allerdings dieser Ansatz wird nicht als eine hilfreiche Reduktion verstanden, sondern tatsächlich mit der Essenz rationalen Denkens identifiziert. In anderen Worten, die Formalisierung der Bedingungsontologie ist selbst *hinreichend* um die Realität zu beschreiben. Gegen diese Auffassung möchte ich hier nur ein kurzes Argument skizzieren, welches sich anhand der hier so wichtigen materialen Implikation zeigen lässt: $p \rightarrow q \equiv \neg p \vee q$. In traditioneller philosophischer Einfallslosigkeit ist auch hier mal wieder die Straße nass:

1. Wenn es regnet, wird die Straße nass.

2. Wenn du Zucker ins Wasser gibst, löst er sich auf.

3. Wenn die Wurzel aus 2 eine rationale Zahl ist, dann ist der Mond aus Käse.

   Im Programm einer Idealsprache wäre nun eine mögliche Interpretation, dass die materiale Implikation von den infinit vielen Fällen realer Konditionale abstrahiert und damit das Wesen konditionalen Schließens erfasst (Gabriel, 2016a, S. 121). Dem Gegenüber steht die reale Modalität von Bedingungssätzen. So gibt es zum Beispiel im Altgriechischen neben dem vertrauten Indikativ und dem Konjunktiv auch noch den Optativ (Gabriel, 2016a, S. 120). Dies führt zu feinen Nuancen in der Verwendung, wie folgende Beispiele zeigen:

4. Wann immer einer dies tut, dann freuen sich die Götter (Generell prospektiv).

5. Wärst Du gerade nicht in Münster, fiele ich Dir in die Arme (Irrealis).

Die entscheidende Beobachtung ist, dass es solche Fälle gibt, die von der formalen Logik nicht erfasst werden.[82] Hier tritt einmal mehr die Bedeutung der Verhältnisse vom

[82]Die ist nur ein exemplarischer Fall von einer Reihe von Argumenten, die das Programm einer Ideal-

Allgemeinem zum Konkreten hervor. Die These lautet, dass das *Allgemeine*, das heißt die logische Form nicht das *Konkrete*, die Praxis des Urteilens in einem konkreten Fall prä-determiniert, sondern dass die Praxis des Urteilens unsere logischen Formen modifiziert und *vice versa* (Gabriel, 2016c, S. 381f.). Das heißt formale Logik und natürliche Sprachverwendung sollten im wissenschaftlichen Kontext im präzisen Sinne des Wortes als interdependent betrachtet werden. Auch die Logik steht nicht in einem *Bottom-up* Verhältnis zur Praxis natürlicher Sprachen, sondern auch hier wird ein komplexes reziprokes Verhältnis wirksam. Die logischen Gesetze des Wahrseins bekommen damit eine *konstitutive Plastizität*. Diese Überlegung macht nur deutlich, dass die Logik nicht (eben sowenig wie eine andere formale Sprache) das Wesen der Rationalität erfasst, sondern Aspekte der Rationalität hervorhebt bei gleichzeitiger Vernachlässigung anderer Aspekte (Gabriel, 2016a, S. 120ff.). Der methodische Reduktionismus ist meines Erachtens ein echter Reduktionismus. Er erfasst *nicht* das *Wesentliche* unter Vernachlässigung dessen, was zu vernachlässigen ist, sondern er *zieht Wesentliches ab*, um Anderes hervortreten zu lassen. In der jüngeren Erkenntnistheorie sind Analysen wie die obige dahingehend konsequent weitergedacht worden, dass mit diesen ein monistisches Verständnis von Logik und Wissen fällt, zugunsten einer Pluralität von Logik*en* und logischer Formen (Gabriel, 2016a, S. 121ff.).

Empfehlung: Die Lehre, die aus diesem Argument für die Gütekriterien mitzunehmen ist, ist, dass die individuellen Fälle $S_x ... S_n$ nicht rigoros unter die formalen Logik zu subsumieren sind. Stattdessen sollte die formale Modellierung konditionaler Zusammenhänge im Rahmen des Forums als eine partiell hilfreiche *Reduktion* kenntlich gemacht werden. Das heißt auch, dass die XAI-Beauftragte und andere Verantwortliche für die reale Modalität der Fälle sensibilisiert werden sollten. Für ein ethische sensibilisiertes ISMS Team gilt es diese heuristische Denkweise gewissermaßen als Tugend zu kultivieren. Dies unterstreicht einmal mehr, dass der formale und technische reproduzierbare Ansatz nur furchtbar werden kann, wenn er komplementär mit den institutionellen Gütekriterien gedacht wird, wodurch gewissermaßen der Starrheit der Formalität in Bürokratie und Logik Leben eingehaucht wird.

sprache unter Druck setzen. Für eine umfassendere Auseinandersetzung sind Wittgensteins philosophische Untersuchungen, sowie Putnams und Quines Arbeiten zu empfehlen.

9 Fazit

Auf die drei Eingangs gestellten Leitfragen wurden in dieser Arbeit drei Antworten gegeben. Die Frage nach den Bedingungen von nicht interpretierbarer KI wurde in dieser Arbeit als epistemisches Risiko behandelt. Als Antwort auf diese Herausforderung wurde das normative Ziel epistemischer Sicherheit als Verstehens- und Vertrauensprozess artikuliert. Es wurden schließlich institutionelle und technische Rahmenbedingungen als Gütekriterien vorgeschlagen und diskutiert, die uns dem Ziel der epistemischen Sicherheit näher bringen sollen.

Informationswissenschaften, Komplexitätsforschung, Ethik, Philosophie, Sicherheit und Recht trafen sich hier zu einem Dialog. Das Ergebnis ist eine Komplexitätsstudie, dessen theoretisches Epizentrum der normative Ausgangspunkt eines Menschenbilds der Freiheit, Autonomie und Würde ist. Dabei ist der Mensch ein Wesen in und als Teil von einer (hyper-)komplexen Wirklichkeit. Zu dieser kann sich der Mensch nicht anders verhalten, als urteilend und handelnd unter den Bedingungen des epistemischen Risikos, aber auch der Möglichkeit von Freiheit und Selbstbestimmung. Es wurden Schritte unternommen, um dieses Menschenbild als epistemischen Prozess mit den rechtlichen und technischen Dimensionen von Sicherheit und künstlicher Intelligenz zu kongruieren. Darauf aufbauend wurden erste Gehversuche im Sinne einer experimentellen Evaluation beschritten. Auf diese aufbauend können Verantwortliche die Gütekriterien unter Zunahme der in der Diskussion besprochenen komplexeren Anforderungen anwenden, testen und erweitern.

Abschließend noch eine Bemerkung im Puncto Realismus. Wir brauchen uns nichts vorzumachen, wir leben nicht in der Welt echter informationeller Selbstbestimmung und Integrität, im Gegenteil wird diese im internationalen Kräftemessen von Geheimdiensten, Staaten, organisierter Kriminalität und Unternehmen stetig unterwandert und steht stärker unter Druck denn je. Sie bleibt aber eine Utopie, die insofern eine konkrete ist, als dass wir ihre ethischen Prämissen, technischen und institutionellen Rahmenbedingungen, wie hier versucht, ausformulieren und diskursiv verteidigen können. Spätestens seit der Aufklärung wissen wir zumindest um die Idee echter Freiheit und Selbstbestimmung, wie sie in den Werken des deutschen Idealismus oder der feministischen Theorie ausformuliert wurden. Und bis dahin ist es noch ein weiter Weg, an dem es sich lohnt zu arbeiten, denn „eine andere Welt ist nicht nur möglich, sie ist im Entstehen. Vielleicht werden viele von uns

nicht mehr hier sein, um sie zu begrüßen, aber an einem ruhigen Tag kann ich, wenn ich sehr genau lausche, ihren Atem hören.“ (Arundhati Roy zitiert nach (Luttmer, 2008, S. 3))

Quellenverzeichnis

Amnesty International. (2020). We sense trouble: Automated mass surveillance and predictive policing in the netherlands. https://www.amnesty.org

Ayer, A. J. (1956). *The problem of knowledge*. Macmillan & Co.

Becker, U. (1996). *Das „Menschenbild des Grundgesetzes“ in der Rechtsprechung des Bundesverfassungsgerichts* (Bd. 708). Duncker & Humblot.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Berghoff, C., Neu, M., & von Twickel, A. (2020). Vulnerabilities of connectionist AI applications: Evaluation and defense. *Frontiers in big data*, *3*. https://doi.org/10.3389/fdata.2020.00023

Berti, L., Giorgi, F., & Kasneci, G. (2025). Emergent Abilities in Large Language Models: A Survey. *Arxiv preprint arxiv:2503.05788*. https://arxiv.org/abs/2503.05788

Beschluss des Bundesverfassungsgerichts: “Volkszählungsurteil” (Recht auf informationelle Selbstbestimmung). Zugriff 25. August 2025 unter https://www.bundesverfassungsgericht.de/SharedDocs/Entscheidungen/DE/1983/12/rs19831215_1bvr020983.html

Boehm, O. (2022). *Radikaler Universalismus: Jenseits von Identität* (M. Adrian, Übers.). Propyläen.

Bourdieu, P. (1984). *Die feinen Unterschiede: Kritik der gesellschaftlichen Urteilskraft* (B. Schwibs & A. Russer, Übers.). Suhrkamp Verlag.

Brandom, R. B. (2019). *A spirit of trust: A reading of hegel's phenomenology*. The Belknap Press of Harvard University Press.

Brennan, T., & Dieterich, W. (2017). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS) [First published 29 November 2017]. In J. P. Singh, D. G. Kroner, J. S. Wormith, S. L. Desmarais & Z. Hamilton (Hrsg.), *Handbook of Recidivism Risk/Needs Assessment Tools*. John Wiley & Sons. https://doi.org/10.1002/9781119184256.ch3

Bröckling, M. (2019, 12. Mai). *Interview: Wie die bayerische Polizei das Predictive Policing nach Deutschland brachte*. netzpolitik.org. https://netzpolitik.org/2019/wie-die-bayerische-polizei-das-predictive-policing-nach-deutschland-brachte/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners [Version 4, submitted 22 Jul 2020]. *Arxiv preprint arxiv:2005.14165*. https://arxiv.org/abs/2005.14165

BSI. (2008). *BSI-Standard 100-2: IT-Grundschutz-Vorgehensweise* [Stand: 08. Mai 2008; IT-Grundschutz-Standard]. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/ITGrundschutzstandards/BSI-Standard_1002.html

BSI. (2021). *Sicherer, robuster und nachvollziehbarer Einsatz von KI: Probleme, Maßnahmen und Handlungsbedarfe*. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/kuenstliche-intelligenz_node.html

BSI. (2022). *Formale Methoden und erklärbare Künstliche Intelligenz: Teilergebnis der Projektforschung TK 23*. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Formale_Methoden_erklaerbare_KI.pdf?__blob=publicationFile&v=3

BSI. (2023). *IT-Grundschutz-Kompendium* [Stand: 1. Februar 2023]. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/IT-GS-Kompendium/IT_Grundschutz_Kompendium_Edition2023.pdf?__blob=publicationFile&v=4

BSI. (2024a). *Deep learning reproducibility and explainable ai (xai): Results of bsi's project research*. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Deep_Learning_Reproducibility_and_Explainable_AI.pdf?__blob=publicationFile&v=5

BSI. (2024b). *Einfluss von KI auf die Cyberbedrohungslandschaft*. Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Einfluss_KI_auf_Cyberbedrohungslage.pdf?__blob=publicationFile&v=2

BSI. (2024c). *Die Lage der IT-Sicherheit in Deutschland 2024* [Bericht über die Cybersicherheitslage in Deutschland]. Bundesamt für Sicherheit in der Informationstechnik

(BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/Lageberichte/Lagebericht2024.html

BSI. (2025). *Generative KI-Modelle: Chancen und Risiken für Industrie und Behörden.* Bundesamt für Sicherheit in der Informationstechnik (BSI). https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf?__blob=publicationFile&v=7

Bundesministerium der Justiz und für Verbraucherschutz. (2018). Bundesdatenschutzgesetz (BDSG 2018) [Aktuelle Fassung]. Verfügbar 6. September 2025 unter https://www.gesetze-im-internet.de/bdsg_2018/BJNR209710017.html

Bundesnetzagentur. (2024a). *KI-Service-Desk.* https://www.bundesnetzagentur.de/DE/Fachthemen/Digitales/KI/start_ki.html

Bundesnetzagentur. (2024b). *Zentrale Anlaufstelle (Single Point of Contact).* https://www.bundesnetzagentur.de/DE/Fachthemen/Digitales/KI/12_Anlaufstelle/artikel.html

Caspar, J. (2023). *Wir Datensklaven: Wege aus der digitalen Ausbeutung – Manifest für mehr Freiheits- und Gleichheitsrechte. Wie wir eine demokratische Digitalisierung und informationelle Integrität erreichen können.* Ullstein Buchverlage.

Coeckelbergh, M. (2019). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics*, *26*, 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Correctiv. (2024). *Diese Falschbehauptungen kursieren zur Europawahl 2024.* CORRECTIV – Recherchen für die Gesellschaft. Zugriff 11. Oktober 2025 unter https://correctiv.org/faktencheck/hintergrund/2024/06/03/diese-falschbehauptungen-kursieren-zur-europawahl-2024/

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press.

Crick, F. (1995). *The astonishing hypothesis: The scientific search for the soul.* Scribner.

Dachwitz, I. (2023). *Microsofts Datenmarktplatz Xandr: Das sind 650.000 Kategorien, in die uns die Online-Werbeindustrie einsortiert.* netzpolitik.org. https://netzpolitik.org/2023/microsofts-datenmarktplatz-xandr-das-sind-650-000-kategorien-in-die-uns-die-online-werbeindustrie-einsortiert/

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In *Parallel problem solving from nature – ppsn xvi* (pp. 448–469). Springer International Publishing. https://doi.org/10.1007/978-3-030-58112-1_31

Deng, Z., Dong, Y., Su, H., & Zhu, J. (2021). Discovering and explaining the representation bottleneck of DNNs. *Arxiv preprint arxiv:2111.06236*. https://arxiv.org/abs/2111.06236

Dennett, D. (1993). *Consciousness explained*. Penguin UK.

Dennett, D. (2017). *From bacteria to Bach and back: The evolution of minds*. W. W. Norton & Company.

Dennett, D. (2018). *Daniel dennett on our consciousness, god and other illusions*. SRF Kultur Sternstunden. Retrieved February 6, 2024, from https://www.youtube.com/watch?v=X0ugfVI7UzQ

Deutscher Ethikrat. (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz: Stellungnahme*. https://www.ethikrat.org/publikationen/stellungnahmen/mensch-und-maschine

Deutschlandfunk. (2022). Prädiktive Privatheit: Wieso wir Datenschutz auch kollektiv denken sollten [Ein Gespräch zum Vortrag von Rainer Mühlhoff]. https://www.deutschlandfunknova.de/beitrag/praedikative-privatheit-wieso-wir-datenschutz-auch-kollektiv-denken-sollten#:~:text=Rainer%20M%C3%BChlhoffs%20Definition%20f%C3%BCr%20pr%C3%A4diktive,Willen%20%C3%BCber%20sie%20vorhergesagt%20werden.

Deutschlandfunk. (2025). *Bundesinnenministerium prüft Einsatz von umstrittener Analyse-Software – Kritik von SPD und Grünen* [Online-Artikel]. https://www.deutschlandfunk.de/bundesinnenministerium-prueft-einsatz-von-umstrittener-analyse-software-kritik-von-spd-und-gruenen-104.html

Di Marino, A., Bevilacqua, V., Ciaramella, A., De Falco, I., & Sannino, G. (2025). Ante-hoc methods for interpretable deep models. *ACM Computing Surveys*, *57*(10). https://doi.org/https://doi.org/10.1145/3728637

D'Ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT Press.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use ai in a responsible way*. Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-30371-6

DIW Berlin. (2014). *Who Cares? Die Bedeutung der informellen Pflege durch Erwerbstätige in Deutschland*. DIW Wochenbericht Nr. 14/2014. Zugriff 11. September 2025 unter https://www.diw.de/de/diw_01.c.458718.de/publikationen/wochenberichte/2014_14_2/who_cares__die_bedeutung_der_informellen_pflege_durch_erwerbstaetige_in_deutschland.html

DIW Berlin. (2020). *Millionärinnen unter dem Mikroskop: Datenlücke bei sehr hoher Einkommensteile – Konzentration höher als bisher ausgewiesen*. DIW Wochenbericht Nr. 29/2020. Zugriff 11. September 2025 unter https://www.diw.de/de/diw_01.c.793802.de/publikationen/wochenberichte/2020_29_1/millionaerinnen_unter_dem_mikroskop__datenluecke_bei_sehr_ho___geschlossen______konzentration_hoeher_als_bisher_ausgewiesen.html

DIW Berlin. (2024). *Ausbau der Pflegeversicherung könnte Gender-Care-Gap in Deutschland reduzieren*. DIW Wochenbericht Nr. 7/2024. Zugriff 11. September 2025 unter https://www.diw.de/de/diw_01.c.892941.de/publikationen/wochenberichte/2024_07_1/ausbau_der_pflegeversicherung_koennte_gender_care_gap_in_deutschland_reduzieren.html

Dreier, H., Epping, V., & Lenz, S. (Hrsg.). (2024). *Grundgesetz Kommentar* (10. Aufl.). Springer-Verlag GmbH.

Du, Z., Zeng, A., Dong, Y., & Tang, J. (2025). Understanding emergent abilities of language models from the loss perspective. *Arxiv preprint arxiv:2403.15796*. https://arxiv.org/abs/2403.15796

Ellis, G. F. R. (2012). Top-down causation and emergence: Some comments on mechanisms. *Interface focus*, *2*(1), 126–140.

Ellis, G. F. R. (2016). *How can physics underlie the mind: Top-down causation in the human context*. Springer.

Epping, V., Lenz, S., & Leydecker, P. (2024). *Grundrechte* (10. Aufl.). Springer-Verlag. https://doi.org/10.1007/978-3-662-68609-6

Ernst, H., Schmidt, J., & Beneken, G. (2023). *Grundkurs Informatik: Grundlagen und Konzepte für die erfolgreiche IT-Praxis – Eine umfassende Einführung* (8. Aufl.). Springer Vieweg. https://doi.org/10.1007/978-3-658-41779-6

Ertel, W. (2025). *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung* (6., aktualisierte Aufl.). Springer Fachmedien Wiesbaden.

Ertel, W. (2026). *Einführung in die Künstliche Intelligenz*. Springer Vieweg.

European Commission. (2024a). *Gestaltung der digitalen Zukunft Europas — KI-Gesetz*. European Commission. https://digital-strategy.ec.europa.eu/de/policies/regulatory-framework-ai

European Commission. (2024b). Künstliche Intelligenz – Fragen und Antworten. https://ec.europa.eu/commission/presscorner/api/files/document/print/de/qanda_21_1683/QANDA_21_1683_DE.pdf

European Commission. (2025). Annex to the communication to the commission – approval of the content of the draft communication from the commission: Commission guidelines on the definition of an artificial intelligence system established by regulation (eu) 2024/1689 (ai act). Retrieved October 29, 2025, from https://ec.europa.eu/newsroom/dae/redirection/document/112455

European Digital Media Observatory. (2024). *Eu-related disinformation keeps growing before the eu parliament elections: Monthly brief no. 36 – edmo fact-checking network*. European Digital Media Observatory. Retrieved June 11, 2025, from https://edmo.eu/wp-content/uploads/2024/06/EDMO-36-Horizontal.pdf

European Union (EU). (2016). Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Text von Bedeutung für den EWR). https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32016R0679

European Union (EU). (2024). Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008, (EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz) (Text von Bedeutung für den EWR). https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32024R1689

Farrell, J., & Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. In M. Armstrong & R. Porter (Hrsg.), *Handbook of*

industrial organization (S. 1967–2072). Elsevier. https://ideas.repec.org/h/eee/indchp/3-31.html

Foucault, M. (1973). *Wahnsinn und Gesellschaft: Eine Geschichte des Wahns im Zeitalter der Vernunft* (U. Köppen, Übers.). Suhrkamp.

Foucault, M. (1974). *Die Ordnung des Diskurses: Inauguralvorlesung am Collège de France, 2. Dezember 1970* (W. Seitter, Übers.). Hanser.

Foucault, M. (1975). *Überwachen und Strafen: Die Geburt des Gefängnisses*. Suhrkamp.

Foucault, M. (1991). *Die Ordnung des Diskurses*. Fischer-Taschenbuch-Verlag.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of consciousness studies*, *23*(11–12), 11–39.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *Acm transactions on information systems (TOIS)*, *14*(3), 330–347.

Fröhlich, G., & Rehbein, B. (Hrsg.). (2014). *Bourdieu Handbuch: Leben – Werk – Wirkung* (Sonderausgabe). J. B. Metzler. https://doi.org/10.1007/978-3-476-01379-8

Gabriel, M. (2014). *An den Grenzen der Erkenntnistheorie: Die notwendige Endlichkeit des objektiven Wissens als Lektion des Skeptizismus* (2., verbesserte und erweiterte Auflage) [Erstauflage 2008]. Verlag Karl Alber.

Gabriel, M. (2015). *Ich ist nicht Gehirn. Philosophie des Geistes für das 21. Jahrhundert*. Ullstein Buchverlage GmbH.

Gabriel, M. (2016a). *Die Erkenntnis der Welt*. Verlag Karl Alber.

Gabriel, M. (2016b). *Sinn und Existenz* [Vortrag im Rahmen der Reihe der Hochschule für Philosophie, München, 30. Mai 2016]. Hochschule für Philosophie München. Zugriff 6. Februar 2024 unter https://www.youtube.com/watch?v=tBYMSesOiXc

Gabriel, M. (2016c). *Sinn und Existenz: Eine realistische Ontologie*. Suhrkamp Verlag.

Gabriel, M. (2020a). *Fiktionen*. Suhrkamp Verlag.

Gabriel, M. (2020b). *Neo-Existentialismus*. Verlag Karl Alber.

Gabriel, M. (2021). *Moralischer Fortschritt in dunklen Zeiten: Universale Werte für das 21. Jahrhundert* (1. Auflage). Ullstein Taschenbuchverlag.

Gille, J., Meineck, S., & Dachwitz, I. (2023). *EU country comparison: How data brokers are screening us*. netzpolitik.org. Verfügbar 9. Februar 2024 unter https://netzpolitik.org/2023/eu-country-comparison-how-data-brokers-are-screening-us/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. https://www.deeplearningbook.org/

Goodman, B., & Flaxman, S. (2016). European union regulations on algorithmic decision-making and a "right to explanation". *Arxiv preprint arxiv:1606.08813.* https://arxiv.org/abs/1606.08813

Graham, G. (2023). *Behaviorism.* Stanford Encyclopedia of Philosophy. Verfügbar 29. Oktober 2025 unter https://plato.stanford.edu/entries/behaviorism/

Grundgesetz für die Bundesrepublik Deutschland, Bonn (1949). Zugriff 28. Oktober 2025 unter https://www.gesetze-im-internet.de/gg/

Guo, H., Tao, C., Huang, Z., & Zou, W. (2025). Coverage-guided testing for deep learning models: A comprehensive survey. https://arxiv.org/abs/2507.00496

Habermas, J. (1992). *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats* (1. Auflage). Suhrkamp.

Halpern, J. Y. (2016). *Actual causality.* The MIT Press. Retrieved October 29, 2025, from http://direct.mit.edu/books/oa-monograph-pdf/2262849/book_9780262336611.pdf

Haugen, F. (2023). *Die Wahrheit über Facebook: Warum ich zur Whistleblowerin wurde und was die größte Social-Media-Plattform der Welt so gefährlich macht. Der Insiderbericht einer mutigen Frau* (K. Petersen & A. Lerz, Übers.). Ullstein Buchverlage.

Heidegger, M. (1954). Die Frage nach der Technik. In *Vorträge und Aufsätze* (S. 13–44). Neske.

Heidegger, M. (1977). *Sein und Zeit.* Vittorio Klostermann. (Original erschienen 1927)

Henning, T. (2016). *Kants Ethik: Eine Einführung.* Reclam.

Herd, B., Mata, N., Zafar, S., Heidemann, L., Kelly, J., Zamanian, A., & Tsai, W.-T. (2024). The european artificial intelligence act: Overview and recommendations for compliance. Retrieved May 14, 2025, from https://publica.fraunhofer.de/entities/publication/f87f2c06-4abc-4b6c-987b-3e3d5413a923

Hochrangige Expertengruppe für Künstliche Intelligenz. (2019). *Ethik-Leitlinien für eine vertrauenswürdige KI* [Eingesetzt von der Europäischen Kommission im Juni 2018]. Europäische Kommission. Brüssel. https://demographie-netzwerk.de/site/assets/files/4421/ethik-leitlinien_fur_eine_vertrauenswurdige_ki_1.pdf

Hogrebe, W. (2006). *Echo des Nichtwissens*. Akademie Verlag.

Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *Plos one*, *13*(7). https://doi.org/10.1371/journal.pone.0201016

Honneth, A. (1992). *Kampf um Anerkennung: Zur moralischen Grammatik sozialer Konflikte* (1. Auflage). Suhrkamp.

Hu, Q., Guo, Y., Xie, X., Cordy, M., Ma, L., Papadakis, M., & Le Traon, Y. (2024). Test optimization in DNN testing: A survey. *Acm transactions on software engineering and methodology*, *33*(4), 1–42. https://doi.org/10.1145/3643678

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., & Yi, X. (2020). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer science review*, *37*. https://doi.org/10.1016/j.cosrev.2020.100270

IGLU/PIRLS Germany. (2025). *IGLU / PIRLS Deutschland: Field Trial zur nächsten Hauptstudie 2026* [Abgerufen aktueller Stand; Hauptstudie geplant für 2026]. https://www.iea.nl/studies/germany/IGLU

Institut zur Qualitätsentwicklung im Bildungswesen (IQB). (2022). *IQB-Bildungstrend 2022: Kompetenzrückgänge in Deutsch, aber weitere Fortschritte in Englisch*. Zugriff 11. Oktober 2025 unter https://www.kmk.org/aktuelles/artikelansicht/iqb-bildungstrend-2022-kompetenzrueckgaenge-in-deutsch-aber-weitere-fortschritte-in-englisch.html

Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern recognition letters*, *150*, 228–234. https://www.sciencedirect.com/science/article/pii/S0167865521002440

Johnston, M. (2009). *Saving god. religion after idolatry*. University Press.

Kaiser, G. (2024). *Anteil der Nutzer von WhatsApp nach Generationen in Deutschland im Jahr 2024*. Statista. https://de.statista.com/statistik/daten/studie/510985/umfrage/anteil-der-nutzer-von-whatsapp-nach-altersgruppen-in-deutschland/

Kaiser, L. (2017). *Düsterer Dienst: Recherche deckt Geschäftspraktiken von Palantir auf*. netzpolitik.org. Zugriff 6. Februar 2024 unter https://netzpolitik.org/2017/duesterer-dienst-recherche-deckt-geschaeftspraktiken-von-palantir-auf/

Kant, I. (1781). *Kritik der reinen Vernunft* (Studienausgabe, 9. Auflage). Felix Meiner Verlag.

Kant, I. (2000). *Kritik der praktischen Vernunft. Grundlegung zur Metaphysik der Sitten: Werkausgabe in 12 Bänden* (W. Weischedel, Hrsg.; Bd. VII). Suhrkamp Verlag.

Kaplan, J., McCandlish, S., Henighan, T., Gray, S., Chess, B., Brown, T. B., Radford, A., Child, R., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *Arxiv preprint arxiv:2001.08361*. https://arxiv.org/abs/2001.08361

Kim, M. (2023). *Your web browsing habits may be less private than you think*. IBM Blog. https://research.ibm.com/blog/browser-fingerprinting

Kirchschläger, P. G. (2022). Ethische KI? Datenbasierte Systeme (DS) mit Ethik. *HMD – Praxis der Wirtschaftsinformatik*, *59*, 482–494. https://doi.org/10.1365/s40702-022-00843-2

Knappik, F. (2013). *Im Reich der Freiheit: Hegels Theorie autonomer Vernunft* (J. Halfwassen, D. Perler & M. Quante, Hrsg.; Bd. 114). Walter de Gruyter. https://www.degruyter.com/document/doi/10.1515/9783110299212/html

Kojima, T., Reid, M., Matsuo, Y., Gu, S. S., & Iwasawa, Y. (2023). Large language models are zero-shot reasoners [Version 4, submitted 29 Jan 2023]. *Arxiv preprint arxiv:2205.11916*. https://arxiv.org/abs/2205.11916

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.

Kreis, G. (2015). *Negative Dialektik des Unendlichen: Kant, Hegel, Cantor* (1. Auflage). Suhrkamp.

Leese, M., & Ugolini, V. (2024). Politics of creep: Latent development, technology monitoring, and the evolution of the Schengen information system. *European journal of international security*, *9*(3), 340–356. https://doi.org/10.1017/eis.2024.5

Leibniz, G. (1998). *Monadologie* (Deutsch-Französische Ausgabe). Suhrkamp.

Leisegang, D. (2024). *EU-Rat: KI-Verordnung erhält grünes Licht*. netzpolitik.org. Zugriff 9. Februar 2024 unter https://netzpolitik.org/2024/eu-rat-ki-verordnung-erhaelt-gruenes-licht/

Lewalter, D., Diedrich, J., Goldhammer, F., Reiss, K., & Köller, O. (Hrsg.). (2023). *PISA 2022: Analyse der Bildungsergebnisse in Deutschland*. Waxmann Verlag. https://doi.org/10.31244/9783830998488

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2024). Emergent world representations: Exploring a sequence model trained on a synthetic task. https://arxiv.org/abs/2210.13382

Liang, H., He, E., Zhao, Y., Jia, Z., & Li, H. (2022). Adversarial attack and defense: A survey. *Electronics*, *11*(8). https://doi.org/10.3390/electronics11081283

Lin, Y.-H., Chiang, C.-L., Lin, P.-H., Chang, L.-R., Ko, C.-H., Lee, Y.-H., & Lin, S.-H. (2016). Proposed diagnostic criteria for smartphone addiction. *PLOS One*, *11*(11). https://doi.org/10.1371/journal.pone.0163010

Linartas, M. (2025). *Unverdiente Ungleichheit: Wie der Weg aus der Erbengesellschaft gelingen kann* (1. Auflage). Rowohlt Verlag.

Lippmann, W. (2018). *Die öffentliche Meinung: Wie sie entsteht und manipuliert wird* (W. Förster, Übers.; 1. Aufl.) [Originaltitel: Public Opinion (1922)]. Westend Verlag.

Lipton, P. (2004). *Inference to the best explanation* (2. Aufl.). Routledge.

Lipton, Z. C. (2016). The mythos of model interpretability. *Arxiv preprint arxiv:1606.03490*. https://arxiv.org/abs/1606.03490

Luhmann, N. (1968). *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität* (4. Auflage). Lucius & Lucius.

Luhmann, N. (1969). *Legitimation durch Verfahren*. Suhrkamp.

Luttmer, M. (2008). *Die AG „Für den Frieden“ und die Sinti und Roma: Versuche aus der Schule zur Unterstützung der Emanzipation einer Minderheit* [Dissertation, Universität Oldenburg]. https://oops.uni-oldenburg.de/2420/1/lutagf08.pdf

Meaker, M. (2023). Slovakia's election deepfakes show ai is a danger to democracy. *Wired*. Retrieved June 11, 2025, from https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

Meineck, S., Köver, C., & Leisegang, D. (2024). *Grundrechte in Gefahr: Die sieben quälendsten Fragen zur KI-Verordnung*. netzpolitik.org. Zugriff 9. Februar 2024 unter https://netzpolitik.org/2024/grundrechte-in-gefahr-die-sieben-quaelendsten-fragen-zur-ki-verordnung/

Meng, M. H., Bai, G., Teo, S. G., Hou, Z., Xiao, Y., Lin, Y., & Dong, J. S. (2022). Adversarial robustness of deep neural networks: A survey from a formal verification

perspective. *Arxiv preprint arxiv:2206.12227*. Retrieved October 29, 2025, from https://arxiv.org/abs/2206.12227

Meyer, U. (2024). *Willensfreiheit, Wissenschaft und diskursive Vernunft: Überlegungen zu Philosophie und (Forschungs-)Praxis*. Brill mentis.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mothilal, R. K., & Tan, C. (2021). Towards unifying feature attribution and counterfactual explanations: different means to the same end. *Arxiv preprint arxiv:2011.04917*. https://arxiv.org/abs/2011.04917

Mühlhoff, R. (2018a). Digitale Entmündigung und User Experience Design. *Leviathan*, *46*(4), 551–574.

Mühlhoff, R. (2018b). *Immersive Macht: Affekttheorie nach Spinoza und Foucault*. Campus Verlag.

Mühlhoff, R. (2020). Automatisierte Ungleichheit: Ethik der künstlichen Intelligenz in der biopolitischen Wende des digitalen Kapitalismus. *Deutsche Zeitschrift für Philosophie*, *68*(6), 867–890.

Mühlhoff, R. (2023a). *Die Macht der Daten: Warum künstliche Intelligenz eine Frage der Ethik ist*. V&R unipress.

Mühlhoff, R. (2023b). KI – Macht – Ungleichheit: Was ist die soziale Dimension von Nachhaltigkeit und warum ist sie durch KI gefährdet? [37C3 – 37th Chaos Communication Congress]. https://media.ccc.de/v/37c3-11937-ki_macht_ungleichheit#t=1216

Mühlhoff, R. (2023c). Predictive privacy: Collective data protection in the context of artificial intelligence and big data. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231166886

Müller, H., & Weichert, F. (2023). *Vorkurs Informatik: Der Einstieg ins Informatikstudium* (6. Aufl.). Springer Vieweg. https://doi.org/10.1007/978-3-658-36468-7

Müller, O., & Lazar, V. (2024). *Transparenz von KI-Systemen* (Whitepaper Version 1.0). Bundesamt für Sicherheit in der Informationstechnik (BSI). Bonn. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper-Transparenz-KI-Systeme.pdf?__blob=publicationFile&v=3

Naarttijärvi, M. (2022). Function creep, altered affordances, and safeguard rollbacks: The many ways to slip on a slippery slope. https://verfassungsblog.de/os6-function-creep/

Nagel, T. (1978). *The possibility of altruism.* Princeton University Press.

Nagel, T. (1986). *The view from nowhere.* Oxford University Press.

Nagel, T. (2016). *Geist und Kosmos: Warum die materialistische neodarwinistische Konzeption der Natur so gut wie sicher falsch ist* (K. Wördemann, Übers.; 5. Aufl.). Suhrkamp Verlag.

Navarro, R., Larrañaga, E., Yubero, S., & Víllora, B. (2020). Psychological Correlates of Ghosting and Breadcrumbing Experiences: A Preliminary Study among Adults. *International journal of environmental research and public health*, *17*(3), 1116. https://doi.org/10.3390/ijerph17031116

Netzpolitik.org. (2024a). *Automatisierte Polizeidatenanalyse: Bayern testet rechtswidrig Palantir-Software.* Zugriff 27. Oktober 2025 unter https://netzpolitik.org/2024/automatisierte-polizeidatenanalyse-bayern-testet-rechtswidrig-palantir-software/

Netzpolitik.org. (2024b). *Digitale Mündigkeit: WhatsApp? Nein danke.* Zugriff 11. September 2025 unter https://netzpolitik.org/2024/digitale-muendigkeit-whatsapp-nein-danke/

Netzpolitik.org. (2025a). *Automatisierte Datenanalyse: Sachsen-Anhalt will interimsweise Palantir.* https://netzpolitik.org/2025/automatisierte-datenanalyse-sachsen-anhalt-will-interimsweise-palantir/

Netzpolitik.org. (2025b). *Verfassungsbeschwerde: Das Problem heißt nicht nur Palantir.* https://netzpolitik.org/2025/verfassungsbeschwerde-das-problem-heisst-nicht-nur-palantir/

Niehus, J. (2025). Technischer Anhang zur Masterarbeit: Gütekriterien sicherer und interpretierbarer Hochrisiko-KI-Systeme [GitHub release]. https://github.com/PyJonny22/Masterarbeit_Guetekriterien-sichere-und-interpetierbare-Hochrisiko-KI-Systeme/releases/tag/v02

Noble, D. (2017). *Dance to the tune of life: Biological relativity.* Cambridge University Press.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism.* NYU Press.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., ... Olah, C. (2022). In-context learning and induction heads. *Anthropic*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Organisation for Economic Co-operation and Development (OECD). (2023). *Pisa 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. Paris. https://doi.org/10.1787/53f23881-en

Pereira, G., & Raetzsch, C. (2022). From banal surveillance to function creep: Automated license plate recognition (ALPR) in Denmark. *Surveillance & Society*, *20*(3), 265–280. https://eprints.lse.ac.uk/116696/

Perlovsky, L. I. (1998). Conundrum of combinatorial complexity. *Proceedings of the ieee international joint conference on neural networks (IJCNN)*, 1464–1469. https://dl.acm.org/doi/abs/10.1109/34.683784

Piketty, T. (2014). *Das Kapital im 21. Jahrhundert*. C. H. Beck.

Popper, K. R. (1934). *Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft*. Mohr Siebeck.

Pörksen, B. (2000). Das Menschenbild der Künstlichen Intelligenz. Ein Gespräch mit Joseph Weizenbaum. *Communicatio Socialis*, *33*(1), 4–17.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *Acm computing surveys (CSUR)*, *51*(5), 1–36.

Qiu, G., Kuang, D., & Goel, S. (2024). Complexity matters: Feature learning in the presence of spurious correlations. *Proceedings of the 41st international conference on machine learning (ICML)*, *235*, 41658–41697. https://proceedings.mlr.press/v235/qiu24e.html

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI technical report*. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* '20: proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Reinhardt, S. (2025). *IT-Grundschutz - Auf dem Weg zum digitalen Regelwerk.* KES Informationssicherheit. Zugriff 27. Oktober 2025 unter https://www.kes-informationssicherheit.de/print/titelthema-metriken-und-kennzahlen/it-grundschutz-auf-dem-weg-zum-digitalen-regelwerk/

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In J. DeNero, M. Finlayson & S. Reddy (Hrsg.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (S. 97–101). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3020

Riddell, R., Ahmed, N., Maitland, A., Lawson, M., & Taneja, A. (2024). Inequality Inc.: How corporate power divides our world and the need for a new era of public action [Briefing Paper]. https://doi.org/10.21201/2024.000007

Ridge, M. (2025). Moral non-naturalism [Abschnitt "The Naturalistic Fallacy"]. In E. Zalta & U. Nodelman (Eds.), *The stanford encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/moral-non-naturalism/

Rudl, T. (2018). *Whistleblower: Überwachungskonzern Palantir hat Cambridge Analytica bei illegalen Methoden geholfen.* netzpolitik.org. Zugriff 6. Februar 2024 unter https://netzpolitik.org/2018/whistleblower-ueberwachungskonzern-palantir-hat-cambridge-analytica-bei-illegalen-methoden-geholfen/

Rumpf, H.-J., Achab, S., Billieux, J., Bowden-Jones, H., Carragher, N., Demetrovics, Z., Higuchi, S., King, D. L., Mann, K., Potenza, M., Saunders, J. B., Abbott, M., Ambekar, A., Aricak, O. T., Assanangkornchai, S., Bahar, N., Borges, G., Brand, M., Chan, E. M.-L., … Poznyak, V. (2018). Including gaming disorder in the icd-11: The need to do so from a clinical and public health perspective. *Journal of behavioral addictions*, *7*(3), 556–561. https://doi.org/10.1556/2006.7.2018.59

Russell, S., & Norvig, P. (2024). *Artificial intelligence: A modern approach* (4. Aufl.). Pearson.

Saleem, R., Yuan, B., Kurugollu, F., Anjum, A., & Liu, L. (2022). Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, *513*, 165–180. https://www.sciencedirect.com/science/article/pii/S0925231222012218?utm_source=chatgpt.com

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press.

Schaffer, J. (2010). Monism: The priority of the whole. *Philosophical review*, *119*(1), 31–76.

Schütze, P. (2022). *Mining the future? The artificial intelligence of climate breakdown* [Beitrag im Blog „Le Club / Berliner Gazette", Teil der Textreihe „After Extractivism"]. Verfügbar 15. Oktober 2025 unter https://blogs.mediapart.fr/berliner-gazette/blog/260522/mining-future-artificial-intelligence-climate-breakdown

Searle, J. (1989). Intentionalität. Eine Abhandlung zur Philosophie des Geistes. *Zeitschrift für philosophische forschung*, *43*(2), 393–397.

Sellars, W. (1999). *Der Empirismus und die Philosophie des Geistes* (T. Blume, Hrsg. & Übers.). Mentis Verlag.

Sen, A. (2006). *Identity and violence: The illusion of destiny.* W. W. Norton & Company.

Sherman, A. (2024). *Fake Joe Biden robocall in new hampshire tells democrats not to vote in the primary election.* Retrieved June 11, 2025, from https://www.politifact.com/factchecks/2024/jan/22/robocaller/fake-joe-biden-robocall-in-new-hampshire-tells-dem/

Stanford Encyclopedia of Philosophy. (2025). *Algorithmic fairness.* Retrieved February 9, 2025, from https://plato.stanford.edu/entries/algorithmic-fairness/

Stephan, A., & Walter, S. (2013). *Handbuch Kognitionswissenschaft.* Springer-Verlag.

The World Bank. (2024). *Pathways out of poverty toward a more prosperous future.* World Bank Blog. https://blogs.worldbank.org/en/voices/pathways-out-of-poverty-toward-a-more-prosperous-future

Tomasello, M. (2020). *Mensch werden. Eine Theorie der Ontogenese* (J. Schröder, Übers.). Suhrkamp Verlag.

Tooze, A. (2022). *Zeitenwende oder Polykrise? Das Modell Deutschland auf dem Prüfstand* [Willy Brandt Lecture 2022, YouTube, Minute 14:00]. Bundeskanzler-Willy-Brandt-Stiftung. https://www.youtube.com/watch?v=K80HOp5MOpA

Tooze, A. (2025). *Chartbook 262 crisis tribes – on europe now* [Substack-Blog]. Retrieved August 25, 2025, from https://adamtooze.substack.com/p/chartbook-262-crisis-tribes-on-europe

Tschantz, M. C. (2022). What is proxy discrimination? *FAccT '22: proceedings of the 2022 acm conference on fairness, accountability, and transparency*, 1993–2003. https://doi.org/10.1145/3531146.3533242

Tu, L., Lalwani, G., Gella, S., & He, H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the association for computational linguistics*, *8*, 621–633. https://doi.org/10.1162/tacl_a_00335

Vankov, I., & Bowers, J. (2019). Training neural networks to encode symbols enables combinatorial generalization. *Arxiv preprint arxiv:1903.12354*. https://arxiv.org/abs/1903.12354

Voosholz, J., & Gabriel, M. (2021). *Top-down causation and emergence*. Springer.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *arXiv preprint arXiv:1711.00399*. https://arxiv.org/abs/1711.00399

Watson, D. S., Gultchin, L., Taly, A., & Floridi, L. (2021). Local explanations via necessity and sufficiency: Unifying theory and practice. *Proceedings of the 35th aaai conference on artificial intelligence (aaai 2021)*, 10323–10331. https://arxiv.org/pdf/2103.14651

Weber, M. (1922). *Wirtschaft und Gesellschaft: Grundriß der verstehenden Soziologie*. Mohr Siebeck.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD

Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136. Retrieved June 10, 2025, from http://www.jstor.org/stable/20024652

World Medical Association. (2024). Wma declaration of helsinki – ethical principles for medical research involving human participants. Retrieved September 11, 2025, from https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Wu, S., Yuksekgonul, M., Zhang, L., & Zou, J. (2023). Discover and cure: Concept-aware mitigation of spurious correlation. *Proceedings of machine learning research.* https://proceedings.mlr.press/v202/wu23w/wu23w.pdf

Ye, W., Jiang, L., Xie, E., Zheng, G., Ma, Y., Cao, X., Guo, D., Qi, D., He, Z., Tian, Y., Coffee, M., Zeng, Z., Li, S., Huang, T.-h. (, Wang, Z., Rehg, J. M., Kautz, H., & Zhang, A. (2024). The clever hans mirage: A comprehensive survey on spurious correlations in machine learning. *Arxiv preprint arxiv:2402.12715.* https://arxiv.org/abs/2402.12715

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *5*(5), 726–742.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* PublicAffairs.

Appendix

9.1 Hochrisikosysteme

Der folgende Text ist ein direkter Auszug aus der EU KI-Verordnung:

Hochrisiko-KI-Systeme gemäß Artikel 6 Absatz 2 (European Union (EU), 2024, Art. 6) Als Hochrisiko-KI-Systeme gemäß Artikel 6 Absatz 2 gelten die in folgenden Bereichen aufgeführten KI-Systeme:

1. Biometrie, soweit ihr Einsatz nach einschlägigem Unionsrecht oder nationalem Recht zugelassen ist:

   (a) biometrische Fernidentifizierungssysteme: Dazu gehören nicht KI-Systeme, die bestimmungsgemäß für die biometrische Verifizierung, deren einziger Zweck darin besteht, zu bestätigen, dass eine bestimmte natürliche Person die Person ist, für die sie sich ausgibt,verwendet werden sollen;

   (b) KI-Systeme, die bestimmungsgemäß für die biometrische Kategorisierung nach sensiblen oder geschützten Attributen oder Merkmalen auf der Grundlage von Rückschlüssen auf diese Attribute oder Merkmale verwendet werden sollen;

   (c) KI-Systeme, die bestimmungsgemäß zur Emotionserkennung verwendet werden sollen.

2. Kritische Infrastruktur: KI-Systeme, die bestimmungsgemäß als Sicherheitsbauteile im Rahmen der Verwaltung und des Betriebs kritischer digitaler Infrastruktur, des Straßenverkehrs oder der Wasser-, Gas-, Wärme- oder Stromversorgung verwendet werden sollen.

3. Allgemeine und berufliche Bildung:

   (a) KI-Systeme, die bestimmungsgemäß zur Feststellung des Zugangs oder der Zulassung oder zur Zuweisung natürlicher Personen zu Einrichtungen aller Ebenen der allgemeinen und beruflichen Bildung verwendet werden sollen;

   (b) KI-Systeme, die bestimmungsgemäß für die Bewertung von Lernergebnissen verwendet werden sollen, einschließlich des Falles, dass diese Ergebnisse dazu die-

nen, den Lernprozess natürlicher Personen in Einrichtungen oder Programmen aller Ebenen der allgemeinen und beruflichen Bildung zu steuern;

(c) KI-Systeme, die bestimmungsgemäß zum Zweck der Bewertung des angemessenen Bildungsniveaus, das eine Person im Rahmen von oder innerhalb von Einrichtungen aller Ebenen der allgemeinen und beruflichen Bildung erhalten wird oder zu denen sie Zugang erhalten wird, verwendet werden sollen;

(d) KI-Systeme, die bestimmungsgemäß zur Überwachung und Erkennung von verbotenem Verhalten von Schülern bei Prüfungen im Rahmen von oder innerhalb von Einrichtungen aller Ebenen der allgemeinen und beruflichen Bildung verwendet werden sollen.

4. Beschäftigung, Personalmanagement und Zugang zur Selbstständigkeit:

(a) KI-Systeme, die bestimmungsgemäß für die Einstellung oder Auswahl natürlicher Personen verwendet werden sollen, insbesondere um gezielte Stellenanzeigen zu schalten, Bewerbungen zu sichten oder zu filtern und Bewerber zu bewerten;

(b) KI-Systeme, die bestimmungsgemäß für Entscheidungen, die die Bedingungen von Arbeitsverhältnissen, Beförderungen und Kündigungen von Arbeitsvertragsverhältnissen beeinflussen, für die Zuweisung von Aufgaben aufgrund des individuellen Verhaltens oder persönlicher Merkmale oder Eigenschaften oder für die Beobachtung und Bewertung der Leistung und des Verhaltens von Personen in solchen Beschäftigungsverhältnissen verwendet werden soll.

5. Zugänglichkeit und Inanspruchnahme grundlegender privater und grundlegender öffentlicher Dienste und Leistungen:

(a) KI-Systeme, die bestimmungsgemäß von Behörden oder im Namen von Behörden verwendet werden sollen, um zu beurteilen, ob natürliche Personen Anspruch auf grundlegende öffentliche Unterstützungsleistungen und -dienste, einschließlich Gesundheitsdiensten, haben und ob solche Leistungen und Dienste zu gewähren, einzuschränken, zu widerrufen oder zurückzufordern sind;

(b) KI-Systeme, die bestimmungsgemäß für die Kreditwürdigkeitsprüfung und Bo-

nitätsbewertung natürlicher Personen verwendet werden sollen, mit Ausnahme von KI-Systemen, die zur Aufdeckung von Finanzbetrug verwendet werden;

(c) KI-Systeme, die bestimmungsgemäß für die Risikobewertung und Preisbildung in Bezug auf natürliche Personen im Fall von Lebens- und Krankenversicherungen verwendet werden sollen;

(d) KI-Systeme, die bestimmungsgemäß zur Bewertung und Klassifizierung von Notrufen von natürlichen Personen oder für die Entsendung oder Priorisierung des Einsatzes von Not- und Rettungsdiensten, einschließlich Polizei, Feuerwehr und medizinischer Nothilfe, sowie für Systeme für die Triage von Patienten bei der Notfallversorgung verwendet werden sollen.

6. Strafverfolgung, soweit ihr Einsatz nach einschlägigem Unionsrecht oder nationalem Recht zugelassen ist:

(a) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden oder in deren Namen zur Bewertung des Risikos einer natürlichen Person, zum Opfer von Straftaten zu werden, verwendet werden sollen;

(b) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden als Lügendetektoren oder ähnliche Instrumente verwendet werden sollen;

(c) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Bewertung der Verlässlichkeit von Beweismitteln im Zuge der Ermittlung oder Verfolgung von Straftaten verwendet werden sollen;

(d) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Bewertung des Risikos, dass

eine natürliche Person eine Straftat begeht oder erneut begeht, nicht nur auf der Grundlage der Erstellung von Profilen natürlicher Personen gemäß Artikel 3 Absatz 4 der Richtlinie (EU) 2016/680 oder zur Bewertung persönlicher Merkmale und Eigenschaften oder vergangenen kriminellen Verhaltens von natürlichen Personen oder Gruppen verwendet werden sollen;

(e) KI-Systeme, die bestimmungsgemäß von Strafverfolgungsbehörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Unterstützung von Strafverfolgungsbehörden zur Erstellung von Profilen natürlicher Personen gemäß Artikel 3 Absatz 4 der Richtlinie (EU) 2016/680 im Zuge der Aufdeckung, Ermittlung oder Verfolgung von Straftaten verwendet werden sollen.

7. Migration, Asyl und Grenzkontrolle, soweit ihr Einsatz nach einschlägigem Unionsrecht oder nationalem Recht zugelassen ist:

(a) KI-Systeme, die bestimmungsgemäß von zuständigen Behörden oder in deren Namen oder Organen, Einrichtungen und sonstigen Stellen der Union als Lügendetektoren verwendet werden sollen oder ähnliche Instrumente;

(b) KI-Systeme, die bestimmungsgemäß von zuständigen Behörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union zur Bewertung eines Risikos verwendet werden sollen, einschließlich eines Sicherheitsrisikos, eines Risikos der irregulären Einwanderung oder eines Gesundheitsrisikos, das von einer natürlichen Person ausgeht, die in das Hoheitsgebiet eines Mitgliedstaats einzureisen beabsichtigt oder eingereist ist;

(c) KI-Systeme, die bestimmungsgemäß von zuständigen Behörden oder in deren Namen oder von Organen, Einrichtungen und sonstigen Stellen der Union verwendet werden sollen, um zuständige Behörden bei der Prüfung von Asyl- und Visumanträgen sowie Aufenthaltstiteln und damit verbundenen Beschwerden im Hinblick auf die Feststellung der Berechtigung der den Antrag stellenden natürlichen Personen, einschließlich damit zusammenhängender Bewertungen der Verlässlichkeit von Beweismitteln, zu unterstützen;

(d) KI-Systeme, die bestimmungsgemäß von oder im Namen der zuständigen Behörden oder Organen, Einrichtungen und sonstigen Stellen der Union, im Zusammenhang mit Migration, Asyl oder Grenzkontrolle zum Zwecke der Aufdeckung, Anerkennung oder Identifizierung natürlicher Personen verwendet werden sollen, mit Ausnahme der Überprüfung von Reisedokumenten.

8. Rechtspflege und demokratische Prozesse:

   (a) KI-Systeme, die bestimmungsgemäß von einer oder im Namen einer Justizbehörde verwendet werden sollen, um eine Justizbehörde bei der Ermittlung und Auslegung von Sachverhalten und Rechtsvorschriften und bei der Anwendung des Rechts auf konkrete Sachverhalte zu unterstützen, oder die auf ähnliche Weise für die alternative Streitbeilegung genutzt werden sollen;

   (b) KI-Systeme, die bestimmungsgemäß verwendet werden sollen, um das Ergebnis einer Wahl oder eines Referendums oder das Wahlverhalten natürlicher Personen bei der Ausübung ihres Wahlrechts bei einer Wahl oder einem Referendum zu beeinflussen. Dazu gehören nicht KI-Systeme, deren Ausgaben natürliche Personen nicht direkt ausgesetzt sind, wie Instrumente zur Organisation, Optimierung oder Strukturierung politischer Kampagnen in administrativer oder logistischer Hinsicht

9.2 Liste von Angriffsvektoren und Beispielen

- Fairness, Diskriminierung und Bias: Diese Risikodimension umfasst die systematische und ungerechtfertigte Ungleichbehandlung von Individuen und Gruppen durch KI-Systeme (Friedman & Nissenbaum, 1996). Dies wird insbesondere relevant, wenn KI-Systeme in Entscheidungen involviert sind, ob Menschen Zugang zu Ressourcen und Chancen bekommen, wie im Falle von Credit-Scoring oder Einstellungsalgorithmen auf dem Arbeitsmarkt (O'Neil, 2017) (Mühlhoff, 2020) (Amnesty International, 2020).

- Qualität und Halluzinieren: KI Modelle erzeugen oft unsinnigen oder falschen Output. Bei dem als Halluzinieren bekanntem Phänomen werden mit einer selbstbewussten Sprache falsche oder qualitativ mangelhafte Inhalte als wahr und hochwertig präsen-

tiert. Dies kann Nutzende dahingehend manipulieren, dass sie von der Korrektheit solcher Ausgaben überzeugt werden (BSI, 2021, 2025).

- Täuschung und Manipulation: Nutzende können mittels (generativer) KI Modelle manipuliert werden. Beispielsweise sind Indirect Prompt Injections Angriffe, bei denen Dritte die Eingabe eines Models gezielt beeinflussen, um das Modellverhalten zu manipulieren. Im Wissen etwa, auf welche Quellen ein Modell bei der Bearbeitung einer Anfrage zugreift, können diese gezielt manipuliert werden, indem Daten hinzugefügt oder entfernt werden (Data Poisoining). Neben den Input- oder Trainingsdaten können auch die Modelle direkt manipuliert werden, indem zum Beispiel die Gewichte angepasst werden (Modell Poisoining). Desweiteren können synthetisch generierte Medien (Deepfakes), von böswilligen Akteuren zu manipulativen Zwecken ausgebeutet werden. Nutzende könnten dann entgegen ihrer Interessen zu Käufen, der Stimmabgabe für eine bestimmte Partei oder zur Preisgabe von Daten motiviert werden. Auch die gezielte, subtile Meinungsmanipulation durch längere Chatverläufe ist denkbar (Berghoff, Neu & von Twickel, 2020; BSI, 2021, 2025).

- Cyberangriffs- und Missbrauchspotenzial: Über die Manipulation der Nutzenden hinaus bestehen noch diverse angrenzende Sicherheitsrisiken. Böswillige Akteure können etwa gezielt Schadcode (Malware) in Bibliotheken platzieren, auf die die Modelle beim Assistieren einer Programmieraufgabe häufig zugreifen oder die sie dem Nutzenden vorschlagen (BSI, 2025). Des Weiteren können durch Remote Code Execution Attacken (RCE) Modelle angeleitet werden, Schadcode zu generieren und diesen auf den Rechner von Nutzenden zu platzieren (BSI, 2024b, 2025).

- Datenschutz und Privatsphäre: Im gesamten Lebenszyklus eines KI-Modells bestehen diverse Risiken für (personenbezogene) Daten und für die Privatsphäre. Beim Bau eines Modells können personenbezogene Daten oder urheberrechtlich geschützte Daten entgegen geltendem Datenschutzrecht für das Training verwendet werden. Diese Verletzung der Rechte von Personen werden dann *post-hoc* durch den Blackbox Charakter des Modells verschleiert. Weiterhin können böswillige Akteure durch Privacy Attacks bzw. Information Extraction Attacks versuchen Trainingsdaten, so auch personenbezogene Daten, zu rekonstruieren (BSI, 2021, 2024b, 2025).

9.3 Komplexität

9.3.1 Beispiele Komplexe Phänomene

Um für den Leser ohne aufwendige Recherche in der zitierten Literatur das abstrakte Modell der Komplexität ein wenig mit Leben zu füllen, seien hier einmal eine Liste von anschaulichen Beispielen aus der Literatur gegeben (Ellis, 2012), (Ellis, 2016), (Voosholz & Gabriel, 2021), (Gabriel, 2020a). Gemein ist diesen, dass sie die beschriebenen Eigenschaften komplexer Systeme erfüllen, das bedeutet wesentlich das hierarchische, strukturierte Zusammenwirken von Top-down und Bottom-up Verursachung, welche zum Entstehen emergenter Phänomene führen:

- **Natürliche Objekte**: Physische Objekte wie Steine, Planeten, Sterne oder Galaxien, bei denen kein Zweck erkennbar ist.
- **Leben**: Zielgerichtete Entitäten wie Bakterien, Pflanzen, Tiere und Menschen.
- **Hergestellte Objekte und gebaute Umwelt**: Artefakte wie Autos, Flugzeuge, Computersysteme, Häuser, Städte, Brücken oder Wassersysteme, die zur Erfüllung bestimmter Zwecke entworfen wurden.
- **Organisationen**: Gesellschaften, Unternehmen, Staaten, Armeen und andere soziale Konstruktionen mit abstrakten und physischen Aspekten.
- **Konzeptuelle Strukturen**: Mentale Strukturen wie Sprache, Mathematik, Theorien, Modelle, Rechtssysteme und Verfassungen.
- **Mathematische Entitäten**: Z. B. die Zahl π, trigonometrische Funktionen oder der Satz des Pythagoras, die technischer Praxis zugrunde liegen.
- **Mentale Auffassungen physikalischer Gesetze**: Etwa Maxwells Gleichungen, die Radio, Radar, Fernsehen oder Mobiltelefone ermöglichen.
- **Computerprogramme und Daten**: Grundlage zahlreicher Anwendungen wie Geldautomaten, Internetbanking, Flugzeugsteuerung oder automatisierte Fabriken.
- **Menschliche Pläne und Intentionen für Alltagsobjekte**: Z. B. Pläne für Computer, Flugzeuge, Flughäfen, Teekannen oder Brillen, die zur Manipulation

vieler Bestandteile führen.

- **Pläne zur experimentellen Manipulation von Mikroentitäten**: Etwa in der Molekülsynthese, Nanotechnologie oder Teilchenerzeugung in Teilchenbeschleunigern (z. B. LHC).
- **Erwartungen und Prognosen**: Zum Beispiel Erwartungen über Preisentwicklungen an Börsen, die reale wirtschaftliche Folgen haben.
- **Soziale Konstruktionen**: Regeln und Systeme wie Schach, Geldwert oder Rechtssysteme, die das gesellschaftliche Zusammenleben ermöglichen.
- **Gesellschaftliche Rollen und Vorbilder**: Rollen wie Lehrer, Richter, Student oder Polizist sowie Vorbilder, die unsere Erwartungen und Handlungen prägen.
- **Information**: Wie durch die Existenz einer IT-Industrie belegt.
- **Schönheit**: Beispielsweise sichtbar daran, dass Häuser mit schöner Aussicht deutlich teurer sind.
- **Sprache**: Ohne die Denken und Intelligenz nicht möglich wären.

9.3.2 Begriffsbestimmung Emergenz

Im folgenden finden Sie eine kurze Bestimmung des Phänomens Emergenz, anhand von vier Kriterien (Ellis, 2016, S. 85ff.):

- Unvorhersagbarkeit. Ein Zustand oder ein Merkmal ist emergent, wenn es entweder prinzipiell oder in der Praxis unmöglich ist, ihn auf der Grundlage einer vollständigen Theorie der grundlegenden Phänomene des Systems vorherzusagen.
- Neue Variablen sind nötig. Man braucht ein neues begriffliches oder beschreibendes Instrumentarium auf höheren Ebenen als das, welches für grundlegendere Phänomene verwendet wird.
- Holismus. Manche Eigenschaften entstehen nur aus Ganzen, die aus der Zusammensetzung grundlegenderer Teile gebildet werden. Es ist begrifflich inkohärent, sie nur in Bezug auf die Teile zu erfassen.

- Das Ganze ist mehr als die Summe seiner Teile. Die Eigenschaften auf Makroebene können nicht durch einfache Addition der Eigenschaften auf niedrigerer Ebene erlangt werden.

9.3.3 5 Formen der Top-down Kausalität

Hier habe ich noch einmal knapp die 5 verschiedenen Formen von Top-down Kausalität nach (Ellis, 2016, S. 133ff.) zusammengefasst. Zum Verständnis der Arbeit ist es nicht notwendig diese genau zu differenzieren, doch ist es hilfreich, zum Verständnis der in der Arbeit selbst nicht verteidigten Prämissen:

Top-down-Kausalität 1: Algorithmische Top-down-Kausalität: Diese Art der Kausalität entsteht, wenn höherstufige Variablen die Dynamik auf niedrigerer Ebene durch die Strukturierung des Systems kausal beeinflussen. Das Ergebnis hängt hierbei eindeutig von den strukturellen, Rand- und Anfangsbedingungen auf höherer Ebene ab, wobei die unteren Ebenen die Ergebnisse algorithmisch ableiten. Beispielsweise bestimmen die höherstufigen Variablen von Computerprogrammen auf der Softwareebene, die Art und Weise, wie sich die Logikgatter auf der Hardwareebene und der Elektronenfluss durch diese verhält.

Top-down-Kausalität 2: Top-down-Kausalität durch nicht-adaptive Informationskontrolle: Hierbei beeinflussen höherrangige Einheiten die niedrigeren Ebenen, um spezifische, feste Ziele zu erreichen, typischerweise durch Regelkreise mit Rückmeldung. Das Ergebnis wird nicht von Anfangs- oder Randbedingungen bestimmt, sondern vom festgelegten Ziel. Ein klassisches Beispiel ist ein Thermostat, das die Raumtemperatur regelt, oder die genetisch festgelegten homöostatischen Systeme in biologischen Organismen.

Top-down-Kausalität 3: Top-down-Kausalität durch adaptive Selektion: Dieser Typus beschreibt Prozesse, bei denen viele Einheiten interagieren und Variationen in ihren Eigenschaften entstehen, gefolgt von der Selektion der am besten an ihre Umgebung angepassten Entitäten. Die höhere Ebene (der Kontext/die ökologische Nische) bietet günstige oder ungünstige Bedingungen, die die Auswahl der unteren Einheiten leiten. Das Ergebnis ist nicht von internen Zielen bestimmt, sondern von Meta-Zielen (Fitnesskriterien), die neue Informationen und Komplexität hervorbringen. Ein Paradebeispiel ist die

Darwinistische Evolution, bei der Umweltbedingungen die Entwicklung von Organismen durch Selektion beeinflussen.

Top-down-Kausalität 4: Top-down-Kausalität durch adaptive Informationskontrolle Diese Form kombiniert nicht-adaptive Informationskontrolle mit adaptiver Selektion der Ziele. Die Ziele des Rückkopplungssystems sind zwar höhere Variablen, aber im Gegensatz zur nicht-adaptiven Kontrolle sind sie nicht fest, sondern können sich adaptiv aufgrund von Erfahrung und Informationen ändern. Dieser Prozess wird durch Fitnesskriterien für die Zielauswahl geleitet und ermöglicht Lernen, Antizipation und flexibles Verhalten. Assoziatives Lernen, wie das Pavlovsche Konditionieren, ist ein Beispiel, bei dem das Gehirn das Verhalten basierend auf Erfahrungen und dem Vermeiden negativer Reize anpasst.

Top-down-Kausalität 5: Intelligente Top-down-Kausalität (d.h. der Einfluss des menschlichen Geistes auf die physische Welt): Dies ist ein Spezialfall der adaptiven Zielwahl, bei dem die Auswahl der Ziele die Verwendung symbolischer Repräsentationen zur Untersuchung von Ergebnissen beinhaltet. Hierbei werden abstrakte, hierarchisch strukturierte symbolische Systeme (wie Sprache, Pläne oder mathematische Modelle) genutzt, um komplexe Situationen zu verstehen, Ergebnisse zu prognostizieren und zukünftige Handlungen rational zu planen. Beispiele sind der Entwurf eines Flugzeugs, bei dem ein abstrakter Plan zur Entstehung eines physischen Objekts führt, oder der Wert von Geld, der auf sozialen Vereinbarungen basiert und physische Veränderungen bewirken kann. Auch Rollen, Erwartungen und Werte in der Gesellschaft üben eine kausale Wirkung von oben nach unten aus. Ellis und Gabriel beanspruchen in der Tat, dass die Konklusion unvermeidlich sei, dass es eine intelligente Top-down Verursachung gibt, bei der abstrakte Entitäten, wie Symbole, Formeln und Gründe in Diskursen einen kausal messbaren Einfluss auf die physische Wirklichkeit haben. „There are various kinds of abstract entities that are causally effective in Top-down causation [...]. They include goals in feedback control systems and selection criteria in adaptive systems. In the case of the mind, they include conscious goals and plans, abstract theories, social constructions, and ethical values.“(Ellis, 2016, S. 204)

Erklärung über den Einsatz von KI-Werkzeugen

Gemäß der Richtlinie „How to responsibly use ChatGPT and other AI tools to support your learning“ der Universität Osnabrück wird im Folgenden aufgeführt, für welche Zwecke KI-gestützte Werkzeuge unterstützend in dieser Arbeit verwendet wurden:

- Ich habe *GitHub Copilot* genutzt, um mich bei der Verbesserung meiner ersten Entwürfe des künstlichen neuronalen Netzes (ANN) für die Evaluation zu unterstützen.
- Ich habe *GitHub Copilot* genutzt, um mich bei der iterativen Verbesserung und Fehlerkorrektur meines Quellcodes für die Evaluation und Evaluation_details zu unterstützen.
- Ich habe *GitHub Copilot* genutzt, um mir einen ersten Entwurf für das Kategorie-Mapping des German Credit Datensatzes zu erstellen. Der Entwurf war fehlerhaft, weswegen ich die Variablen manuell beschriftet habe.
- Ich habe *GitHub Copilot* genutzt, um beim Aufbau und bei der Strukturierung des GitHub-Repositoriums assistiert zu werden.
- Ich habe *GitHub Copilot* genutzt, um Vorschläge für die Kommentierung des Quellcodes zu erhalten. Die Kommentare habe ich jedoch überwiegend selbst verfasst.
- Ich habe GitHub Copilot genutzt, um mir bei der Verbesserung meiner Grafik für die Standardarchitektur zu helfen, die ich mit matplotlib erstellt habe
- Ich habe *ChatGPT* genutzt, um die Syntax von LaTeX (Overleaf) zu erlernen sowie Formeln, Tabellen und Medien korrekt zu formatieren.
- Ich habe *ChatGPT* genutzt, um Vorschläge für die Strukturierung einzelner Abschnitte zu erhalten.
- Ich habe *ChatGPT* bei der Quellenrecherche eingesetzt, um mir einen Überblick über relevante Quellen zu verschaffen und erste inhaltliche Zusammenfassungen zu erstellen.
- Ich habe *ChatGPT* genutzt, um den Text auf Rechtschreibungs- und Grammatikfehler zu prüfen sowie stilistische Verbesserungen vorzuschlagen.

Eigenständigkeitserklärung / Declaration of Authorship

Name, Vorname (Druckbuchstaben) / Full Name (block letters)

Matrikelnummer / Student number

Die Inhalte der hier vorgelegte Arbeit geben meinen eigenen Wissensstand, mein eigenes Verständnis und meine eigene Auffassung zum bearbeiteten Thema wieder. Falls KI-Tools eingesetzt wurden, habe ich deren Einsatzweise und -zweck transparent angegeben. Darüber hinaus habe ich alle meine Quellen akademischen Standards entsprechend ausgewiesen. Ich bin bereit und fähig, die hier erläuterten Inhalte zu erklären und die entwickelten Standpunkte zu vertreten. Die vorliegende Leistung wurde weder zum Teil noch vollständig an dieser oder einer anderen Universität eingereicht.

The content of this thesis represents my own knowledge, my own understanding and my own perspective on the topic. In case artificial intelligence tools were used, their way and purpose of usage has been made transparent. Moreover, I have cited all my sources in accordance with academic standards. I am ready and able to explain and defend the positions developed in this thesis. This thesis has not been submitted, either in part or whole, at this or any other university.

Datum und Unterschrift / Date and Signature