

Technische Universität Dortmund
Fakultät für Humanwissenschaften und Theologie
Institut für Philosophie und Politikwissenschaft

**Dissertation zur Erlangung des akademischen Grades Dr. phil.
im Fach Philosophie**

KI-Systeme instrumentell verstehen

Erklärbarkeit, gesellschaftliche Desiderata und epistemische Opazität

vorgelegt von Sara Mann

Datum der mündlichen Prüfung: 15. Juli 2026

Erstgutachterin: Prof. Dr. Eva Schmidt, TU Dortmund
Zweitgutachterin: Prof. Dr. Lena Kästner, Universität Bayreuth

Veröffentlichung als Dissertation in der Fakultät Humanwissenschaften und Theologie der
Technischen Universität Dortmund im Jahr 2026

Le vrai est trop simple, il faut y arriver
toujours par le compliqué.

George Sand

Danksagung

Im Verlauf der letzten Jahre habe ich, etwa in Kennenlernsituationen, viele Male erzählt, dass ich im Fach Philosophie zu Künstlicher Intelligenz promoviere. Meist reagierte mein Gegenüber mit betretenem Schweigen oder der Frage, was man denn mit einem Philosophie-Doktor mache; manchmal auch mit Lobeshymnen auf ChatGPT oder der Befürchtung, dass wir es bald mit einer allmächtigen Super-KI zu tun bekommen. Besonders im Gedächtnis blieb mir aber eine andere Reaktion: Als ich erwähnte, dass ich promoviere, zögerte meine Gesprächspartnerin keine Sekunde, nahm mich fest in den Arm und versicherte mir mit Nachdruck, dass alles gut werden würde. Besagte Person behielt Recht, und das habe ich einer Vielzahl von Menschen zu verdanken, die mich in den letzten fünf Jahren auf unterschiedlichste, oft supererogative Weise unterstützt haben.

Besonderer Dank gebührt Eva Schmidt, die immer an mich geglaubt und die Arbeit an dieser Dissertation mit konstruktiver Kritik, guten Ideen und der nötigen Prise Pragmatismus begleitet hat. Ich hätte mir keine bessere Betreuerin und Vorgesetzte wünschen können. Ein großes Dankeschön geht außerdem an Lena Kästner, die sich immer wieder Zeit für mich genommen, mir wertvolles Feedback gegeben und neue Blickwinkel aufgezeigt hat – und das, obwohl sie ‚nur‘ meine Zweitgutachterin war.

Ich hatte das große Glück, im Rahmen eines wunderbaren Forschungsprojekts promovieren zu dürfen. Ich bedanke mich von Herzen bei allen Mitgliedern des EIS-Projekts, die das vielbeschworene Schlagwort ‚Interdisziplinarität‘ mit Leben füllten und mir mit viel Humor und geballter Expertise zur Seite standen. Besonderer Dank geht an die ‚Brainy Five‘, insbesondere an Barney Crook, Astrid Schomäcker und Timo Speith, von denen ich nicht zuletzt dank eines unerwartet langwierigen Paper-Projekts viel über das akademische Schreiben und Denken gelernt habe.

Neben den Mitgliedern des EIS-Projekts wurde ich auch durch die Angehörigen des Philosophie-Instituts der TU Dortmund und in der Endphase durch jene des Lamarr-Instituts unterstützt, die sich bei diversen Kolloquien, Mittagspausen und Workshop-Dinners als wertvolle Diskussionspartner:innen bewiesen. Insbesondere möchte ich mich bei Jessica Struchhold bedanken, die mir eine wunderbare Gesprächspartnerin und Freundin war, und bei Swaantje Siebke, deren unermüdliche Unterstützung bei organisatorischen Fragen mir einen freien Kopf fürs Forschen ermöglicht hat.

Viele Inhalte dieser Arbeit wurden in ähnlicher Form auf Workshops, Konferenzen und in Kolloquien vorgestellt und teils in Form von Fachaufsätzen (anonym) begutachtet. Dank gebührt den vielen daran beteiligten Wissenschaftler:innen, deren kritische Rückfragen, Erläuterungen und weiterführende Ideen die hier dargelegten Überlegungen maßgeblich

beeinflusst und geschärft haben. In dieser Hinsicht geht ein besonders großes Dankeschön an die Co-Autor:innen jener Aufsätze, die in abgewandelter Form Teil dieser Arbeit wurden: Barney Crook, Lena Kästner, Markus Langer, Astrid Schomäcker und Timo Speith. Weitere Personen, die mir zu einzelnen Kapiteln oder Aspekten der Arbeit wertvolles Feedback gegeben haben, sind an jeweils passender Stelle in Fußnoten erwähnt.

Bedanken möchte ich mich auch bei den Mitgliedern meiner Doktorand:innen-Schreibgruppe, insbesondere bei Julian Welz, ohne die ich es wohl nie so früh morgens an den Schreibtisch geschafft hätte, sowie bei Andreas Streinzer für den perfekten Dissertations-Workshop zur perfekten Zeit.

Ich war während der gesamten Promotionsphase finanziell abgesichert. Dieses Privileg habe ich dem Einsatz von Lena Kästner, Jakob Rehof und Eva Schmidt zu verdanken, die es mir ermöglichten, die Arbeit an meiner Dissertation nicht nur ohne großen Zeitdruck, sondern auch in einem inspirierenden Arbeitsumfeld zu beenden.

Schließlich gebührt ein riesiger Dank meinem persönlichen Umfeld. Meiner famosen Familie möchte ich für ihre Gelassenheit, ihr Verständnis und die zahllosen Care-Pakete danken, die immer genau im richtigen Moment bei mir eintrafen. Vielen Dank an meine Freund:innen: Jenen, die ebenfalls promovier(t)en, danke ich für den Austausch über die schönen und anstrengenden Seiten des Promovierens – und weil sie mir das ermutigende Gefühl gaben, dass das Schreiben einer Doktorarbeit so wild nun auch wieder nicht ist. Jenen Freund:innen, die sich nicht in dieser speziellen akademischen Blase bewegen, danke ich dafür, dass sie mich immer wieder an das Leben jenseits von Universität und Doktorarbeit erinnerten – und weil sie mir das tröstende Gefühl gaben, dass das Schreiben einer Dissertation halt doch ein heftiges Unterfangen ist. Besonderer Dank geht an Andreas Ketelaer, dessen Informatik-Perspektive – und Python-Skills – auch inhaltlich zu dieser Arbeit beigetragen haben. Vor allem bin ich ihm aber unendlich dankbar für seinen emotionalen Rückhalt, Humor und die unzähligen warmen Mahlzeiten während dieser intensiven und aufregenden Lebensphase.

Finanzielle Förderung Die Arbeit an dieser Dissertation wurde finanziert durch das von der VolkswagenStiftung geförderte EIS-Projekt („Explainable Intelligent Systems“, AZ 98510) sowie durch das Projekt „Explainability, Data Sovereignty, and Trust“ des Dortmunder Lamarr-Instituts für Maschinelles Lernen und Künstliche Intelligenz, gefördert aus Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt und des Landes Nordrhein-Westfalen.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Forschungsfragen	4
1.2	Aufbau der Arbeit	6
I	Grundlagen	9
2	Künstliche Intelligenz	11
2.1	Symbolische versus subsymbolische Ansätze	13
2.1.1	Symbolische KI	13
2.1.2	Subsymbolische KI	14
2.2	Konventionelle KI versus Maschinelles Lernen	17
2.2.1	Konventionelle KI	18
2.2.2	Maschinelles Lernen	19
2.3	Quellen der Opazität von KI-Systemen	21
2.3.1	Architektonische Opazität	23
2.3.2	Analytische Opazität	25
2.3.3	Soziotechnische Opazität	27
3	Erklärbarkeit	29
3.1	Was ist Erklärbarkeit?	29
3.2	Informatische Erklärbarkeitsmethoden	32
3.2.1	Zeitpunkt: ante-hoc versus post-hoc	33
3.2.2	Geltungsbereich: global versus lokal	37
3.2.3	Anwendbarkeit: systemagnostisch versus systemspezifisch	39
3.2.4	Evaluationskriterien für XAI-Erklärungen	40
3.3	KI-Systeme erklären oder verstehen?	42
3.3.1	Sind KI-Systeme erklärbar?	43
3.3.2	Sind Erklärungen notwendig für das Verstehen von KI-Systemen?	52

II	Instrumentelles Verstehen	61
4	Methodik	63
4.1	Begriffsgestaltung und Explikation	63
4.1.1	Warum Begriffsgestaltung?	64
4.1.2	Warum Explikation?	67
4.1.3	Das Rezept	69
4.2	Die Explikationsaufgabe	70
4.2.1	Präzisierung des Explikandums	70
4.2.2	Der theoretische Rahmen des Explikats	74
4.2.3	Die theoretischen Zwecke des Explikats	75
4.3	Adäquatheitsbedingungen des Explikats	80
4.3.1	Ähnlichkeitsbedingungen	80
4.3.2	Nützlichkeitsbedingungen	82
5	Verstehensgrade	85
5.1	Dimensionen der Gradierbarkeit von Verstehen	86
5.1.1	Dimension 1: Vollständigkeit	86
5.1.2	Dimension 2: Faktizität	88
5.1.3	Dimension 3: Begreifen	90
5.1.4	Dimension 4: Rechtfertigung	92
5.1.5	Überschneidungen zwischen den einzelnen Dimensionen	93
5.2	Instrumentelles Verstehen ist nicht gradierbar	95
6	Objekte des Verstehens	97
6.1	Propositionales, interrogatives und objektuales Verstehen	99
6.1.1	Verstehensobjekte als grammatikalische Objekte	100
6.1.2	Verstehensobjekte als Propositionen oder Dinge	101
6.1.3	Verstehensobjekte und Komplexitätsgrade	104
6.2	Symbolisches und faktisches Verstehen	107
6.3	Die Objekte instrumentellen Verstehens sind komplex	109
7	Verstehen und Überzeugung	115
7.1	Verstehen einer Theorie	116
7.2	Verstehen mithilfe einer Theorie	120
7.3	Verstehen nicht-propositionaler Repräsentationen	125
7.3.1	Lösungsstrategien für das Problem der Propositionalität	126
7.3.2	Das Problem der Extraktion	131
7.4	Instrumentelles Verstehen erfordert Überzeugung	132

8 Faktizität von Verstehen	135
8.1 Nicht-faktive Positionen	135
8.2 Faktive Positionen	137
8.3 Moderat faktive Positionen	139
8.3.1 Moderate Faktizität und symbolisches Verstehen	142
8.3.2 Sind falsche Überzeugungen Teil des Verstehens?	144
8.3.3 Zentrale und periphere Überzeugungen	146
8.4 Instrumentelles Verstehen ist moderat faktiv	151
9 Verstehen und Fähigkeiten	157
9.1 Zum Verhältnis von Verstehen und Fähigkeiten	159
9.1.1 Sind Fähigkeiten hinreichend für Verstehen?	159
9.1.2 Sind Fähigkeiten notwendig für Verstehen?	160
9.2 Sechs Cluster verstehensbezogener Fähigkeiten	161
9.2.1 Vorteile verstehensbezogener Fähigkeiten für XAI-Kontexte	163
9.2.2 Cluster 1: erkennen	165
9.2.3 Cluster 2: bewerten	167
9.2.4 Cluster 3: erklären	168
9.2.5 Cluster 4: schlussfolgern	170
9.2.6 Cluster 5: intervenieren	172
9.2.7 Cluster 6: entwickeln	173
9.3 Instrumentelles Verstehen erfordert Fähigkeiten	175
10 Verstehen und Rechtfertigung	181
10.1 Verstehen ohne Rechtfertigung?	182
10.2 Welche Rechtfertigungstheorie eignet sich für Verstehen?	187
10.2.1 Grundfragen der Theorie der Rechtfertigung	187
10.2.2 Begreifen und Rechtfertigung	189
10.2.3 Plädoyer für einen hybriden internalistischen Ansatz	191
10.3 Instrumentelles Verstehen erfordert Rechtfertigung	195
11 Instrumentelles Verstehen	201
11.1 Einführung des Explikats	201
11.1.1 Die Definition	203
11.1.2 Anwendung auf Beispiele	208
11.1.3 Zum Verhältnis von Wissen und (instrumentellem) Verstehen	215
11.2 Evaluation des Explikats	219
11.2.1 Evaluation der Ähnlichkeit	219
11.2.2 Evaluation der Nützlichkeit	222

III Verstehensermöglicher	227
12 Instrumentelles Verstehen und Exemplifikation	229
12.1 Zu Elgins Theorie der Exemplifikation	232
12.2 SAMPLES, EXEMPLARE und instrumentelles Verstehen	237
12.2.1 Eigenschaften von EXEMPLAREN	241
12.2.2 Instrumentelles Verstehen durch EXEMPLARE	245
12.3 Generieren XAI-Methoden EXEMPLARE?	247
12.3.1 Gruppe 1: Ansätze, die Instanzen einer Bildklasse anführen	249
12.3.2 Gruppe 2: Ansätze, die einflussreiche Bereiche in einer Instanz der Bildklasse hervorheben	251
12.3.3 Gruppe 3: Ansätze, die ein künstliches Bild generieren	253
12.4 Einwände	254
12.4.1 Einwand 1: Bildklassifikationssysteme stützen sich auf fremdartige Merkmale	255
12.4.2 Einwand 2: Bildklassifikationssysteme stützen sich auf abstrakte Merkmale	256
12.4.3 Einwand 3: Exemplifikation setzt Verstehen voraus	257
13 Konklusion	259
IV Appendix	267
Abbildungsverzeichnis	269
Tabellenverzeichnis	270
Ergänzende Angaben	271
Literatur	274

Formalia

Vorab einige formale Anmerkungen:

- Die deutsche (und auch englische) Sprache kennt oftmals keine Ausdrücke, um relevante Aspekte von KI-Systemen adäquat zu bezeichnen. Aus Gründen der sprachlichen Einfachheit und Verständlichkeit nutze ich daher mitunter Ausdrücke, die suggerieren, dass es sich bei KI-Systemen um bewusste Entitäten oder Lebewesen handelt (etwa, wenn ich von den Gründen für eine KI-Ausgabe oder den Zielen des Systems spreche). Mein Vorschlag setzt nicht voraus, dass KI-Systeme ein Bewusstsein, genuine Ziele, Wünsche oder eine Gedankenwelt besitzen.
- Ich gendere mit dem Gender-Doppelpunkt (wie in ‚Informatiker:in‘) und weiche punktuell auf das generische Femininum aus. Wenn von Schwarzen Personen die Rede ist, schreibe ich ‚Schwarz‘ groß, um hervorzuheben, dass der Ausdruck keine biologische Eigenschaft, sondern die gesellschaftliche Position von rassifizierten Personen beschreibt.
- Wenn ich mich auf einen Begriff beziehe, ist der korrespondierende Ausdruck *kursiv* gedruckt. Beziehe ich mich auf den zugehörigen Ausdruck, verwende ich einfache Anführungszeichen: ‚ ‚. Wenn aus dem Kontext klar wird, ob ich mich auf den Begriff oder Ausdruck beziehe, verzichte ich mitunter auf diese Hervorhebung. Darüber hinaus dient die *Kursivierung* auch dem Verweis auf Fallbeispiele sowie der sprachlichen Betonung einzelner Wörter bzw. der Hervorhebung fremdsprachiger Ausdrücke; einfache Anführungszeichen (‚ ‚) markieren auch die Ausgaben von KI-Systemen (z. B. die Bildklasse ‚Hund‘) sowie gelegentlich die Verwendung eines Ausdrucks im übertragenen Sinn.
- Interne Verweise auf Kapitel, Abbildungen etc. innerhalb der Arbeit sind ebenso wie Hyperlinks zu Internetadressen serifenlos gedruckt.
- Direkte Zitate sind in doppelte Anführungszeichen („“) eingefasst oder als Blockzitat vom Text abgesetzt. Auslassungen bzw. Einfügungen sind durch eckige Klammern markiert: []
- Besitzen mehrere von mir zitierte Autor:innen denselben Nachnamen, sind in der Kurzzitation im Text die Initialen des Vornamens ergänzt, um eine eindeutige Zuordnung zu gewährleisten (z. B. ‚A. Müller‘).

1 Einleitung

Computersysteme, die auf Künstlicher Intelligenz (KI) beruhen, sind längst selbstverständlicher Bestandteil diverser Lebensbereiche. Mittlerweile werden KI-Systeme genutzt, um Lungenentzündungen oder Hautkrebs auf Basis medizinischer Bilddaten zu diagnostizieren, die Informationsverarbeitung im menschlichen Gehirn zu modellieren, diverse Module in autonomen Fahrzeugen zu steuern, Schädlinge und Krankheiten in der Präzisionslandwirtschaft zu erkennen, Videospiele zu meistern, neue Exoplaneten zu entdecken, visuelle Effekte in Filmen zu produzieren, Klickbetrug und Spamnachrichten zu identifizieren oder vielversprechende Jobkandidat:innen auszuwählen. Und die gegenwärtige Generation großer Sprachmodelle (engl. *large language models*, LLM) und der darauf aufbauenden multimodalen Systeme kann neben Bildern, Videos und Audiomaterial fast jede denkbare Textform in dutzenden von Sprachen, einschließlich Programmiersprachen, generieren. Der Erfolg solcher ‚KI-Systeme mit allgemeinem Verwendungszweck‘ (engl. *general purpose AI*) lässt vermuten, dass die Anwendungsmöglichkeiten von KI in Zukunft noch weiter zunehmen werden.

Gleichzeitig zeigen Forschungsergebnisse und zahllose Skandale, dass KI-Technologie mitunter spektakulär versagt. Manche dieser Fehler haben einen gewissen Unterhaltungswert: Im Jahr 2024 gingen Screenshots von Googles ‚AI Overview‘ viral, also einer KI-generierten Zusammenfassung relevanter Suchergebnisse. Diese Zusammenfassung empfahl unter anderem, jeden Tag einen kleinen Stein zu essen, um den eigenen Mineralstoffbedarf zu decken,¹ oder umgerechnet etwa dreißig Milliliter Klebstoff unter die Pizzasohle zu mischen, um den Käse auf der Pizza zu fixieren.² Und OpenAIs Sprachmodell GPT-5 erntete im August 2025 öffentliche Häme, weil es bei seiner Veröffentlichung die Beine von Tieren auf manipulierten Fotos nicht zählen konnte und z. B. einem offensichtlich fünfbeinigen Zebra lediglich vier Beine bescheinigte.³

Doch ist längst nicht jedes Scheitern eines KI-Systems unterhaltsam. Unter anderem wird immer wieder aufgedeckt, dass KI-Systeme problematische Proxy-Merkmale für die Generierung ihrer Ausgaben heranziehen (Geirhos et al., 2020; Lapuschkin et al., 2019). Beispielsweise konnten Julia Winkler et al. zeigen, dass ein Klassifikationssystem zur Erkennung von Hautkrebs auch auf die Stiftmarkierungen von Dermatolog:innen achtete und

¹Typischerweise werden kuriose Fehler kommerzieller KI-Systeme über soziale Medien verbreitet. Die Empfehlung, Steine zu essen, wurde unter anderem in diesem X-Post festgehalten: <https://x.com/icreatelif/status/1793781850923823144> (besucht am 29.01.2026).

²Siehe etwa <https://x.com/peteryang/status/1793480607198323196> (besucht am 29.01.2026).

³Siehe https://x.com/anh_ng8/status/1953560741367197762 (besucht am 29.01.2026).

nicht bloß auf das Muttermal (Winkler et al., 2019). Da Stiftmarkierungen keinen zuverlässigen Indikator für Melanome darstellen, könnte durch ein solches System schlimmstenfalls eine (potenziell lebensbedrohliche) Hautkrebserkrankung unentdeckt bleiben. Ein anderes Beispiel: Im Jahr 2019 legten Ziad Obermeyer et al. in einem *Science*-Artikel offen, dass ein im US-Gesundheitssystem verbreiteter Algorithmus für die Zuweisung einer gezielten Hochrisiko-Gesundheitsversorgung Schwarze Menschen systematisch benachteiligte. Der Grund: Das System sagte die anfallenden Behandlungskosten vorher, nicht die Erkrankung. Da Schwarze Personen in den USA häufig schlechteren Zugang zu medizinischer Versorgung haben, wies das System ihnen geringere Behandlungskosten und damit auch ein geringeres Risiko zu (Obermeyer et al., 2019). Dadurch reproduzierte das System bestehende systematische Benachteiligungen in der US-amerikanischen Gesellschaft und begünstigte eine ungerechte Verteilung gesundheitsbezogener Ressourcen. All diese Beispiele zeigen, dass moderne KI-Systeme keineswegs fehlerfrei funktionieren – mit potenziell desaströsen Auswirkungen.

Wenn KI-Systeme versagen, werden sie diversen legitimen Anforderungen, Wünschen, Interessen, Zielen und Bedürfnissen nicht gerecht, die Angehörige unterschiedlicher Interessengruppen – so genannte Stakeholder – hinsichtlich KI-Systemen und den übergeordneten soziotechnischen Systemen⁴ haben, in die KI eingebettet ist. Diese Anforderungen sind sehr vielfältig – sie können im Rahmen der Entwicklung, Regulierung oder Verwendung von KI-Systemen bedeutsam sein und von unterschiedlichen Stakeholdern an KI-Kontexte herangetragen werden. Beispielsweise soll ein KI-System, das in der medizinischen Diagnostik eingesetzt wird, vertrauenswürdig sein; ein Empfehlungssystem, das für eine Personalmanagerin die vielversprechendsten Bewerbungen auswählt, sollte fair sein; und der Empfehlungsalgorithmus eines Musikstreamingdienstes sollte genau sein. Im Anschluss an Markus Langer et al. können wir diese Anforderungen als ‚gesellschaftliche Desiderata‘ bezeichnen (Langer et al., 2021).

Leider wird es zunehmend herausfordernder zu gewährleisten, dass KI-Systeme bzw. die übergeordneten soziotechnischen Systeme relevanten Desiderata gerecht werden. Denn viele gegenwärtige KI-Systeme sind zu unterschiedlichen Graden epistemisch opak, d. h. es ist schwierig oder in der Praxis sogar unmöglich, diese Systeme vollständig zu verstehen (Burrell, 2016; Facchini & Termine, 2022). Diese Opazität kann von unterschiedlichen Quellen herrühren. In bestimmten Bereichen (etwa der Bild- oder Sprachverarbeitung) geht der Trend seit Jahren zu immer größeren, komplexeren Systemen, deren Ausgaben selbst für

⁴Der Begriff des soziotechnischen Systems beschreibt (relativ zu einer organisatorischen Einheit) die Gesamtheit von technischen Artefakten sowie sozialen Akteur:innen, deren Interaktionen untereinander und mit den Artefakten. Entsprechend besteht ein soziotechnisches System aus sich überlappenden sozialen sowie technischen Subsystemen und besitzt materielle, soziale und intentionale Elemente, die sich gegenseitig beeinflussen (Leonardi, 2012). Das soziotechnische System, in das ein KI-System eingebettet ist, schließt dieser Perspektive zufolge unter anderem verschiedene Stakeholder, deren Interessen, Anforderungen usw. (‚Desiderata‘ – siehe unten), Arbeits- bzw. Informationsaustauschprozesse, unterschiedliche Verstehensermöglicher (mehr dazu unten), das KI-System selbst sowie kulturelle bzw. soziale Normen ein.

Expert:innen kaum noch nachvollziehbar sind. In anderen Fällen sind KI-Systeme vornehmlich deswegen opak, weil der Hersteller ihre Funktionsweise aus Wettbewerbsgründen geheim hält. Epistemische Opazität erschwert wiederum das Auffinden von Fehlern oder Verzerrungen, die schlimmstenfalls Menschen schaden können.

Zusammengenommen beschreiben diese Überlegungen das so genannte Black-Box-Problem: Einerseits bildet die Leistungsfähigkeit moderner KI-Systeme einen starken Anreiz für ihren Einsatz in diversen, auch risikobehafteten Bereichen. Andererseits gefährdet das unzureichende Verständnis dieser nichtsdestotrotz fehleranfälligen Systeme die Erfüllung gesellschaftlicher Desiderata von relevanten Stakeholdern.

Das Black-Box-Problem hat angesichts mehrerer jüngst verabschiedeter Gesetze zur Regulierung von KI an Dringlichkeit gewonnen. Beispielsweise soll die 2024 in Kraft getretene ‚Verordnung über künstliche Intelligenz‘ der Europäischen Union (engl. *AI Act*; Verordnung (EU) 2024/1689) zum Schutz von Gütern wie Gesundheit, Sicherheit und Grundrechten beim Einsatz von KI beitragen und sicherstellen, dass die in der Europäischen Union verwendeten KI-Systeme vertrauenswürdig sind. Und die Datenschutz-Grundverordnung (DSGVO, Verordnung (EU) 2016/679) gewährt Europäer:innen, deren personenbezogene Daten verarbeitet werden, nach Ansicht einiger Rechtswissenschaftler:innen ein ‚Recht auf Erklärung‘ bei algorithmisch gestützten Entscheidungen (B. Goodman & Flaxman, 2017). Ist ein KI-System jedoch epistemisch opak, erschwert das die Erklärung seiner Ausgaben und die Beurteilung seiner Vertrauenswürdigkeit.

Eine mögliche Lösung für das Black-Box-Problem verspricht das Forschungsfeld der erklärbaren KI (engl. *explainable AI*, kurz: XAI). Die XAI-Forschung bemüht sich darum, relevante Aspekte von KI-Systemen für unterschiedliche Stakeholder verständlich zu machen, indem sie eine Reihe von Erklärbarkeitsmethoden entwickelt (einen Überblick bieten etwa Arrieta et al., 2020; Guidotti et al., 2019; Molnar, 2022). Diese Methoden sollen die Funktionsweise und Entscheidungsprozesse verschiedener Arten von KI-Systemen erklären und dadurch zur Erfüllung diverser Desiderata – wie Fairness, Vertrauenswürdigkeit, menschliche Kontrolle oder Robustheit – beitragen. Zwar macht die Verfügbarkeit einer verständlichen Erklärung ein KI-System weder fair, noch robust, noch genau. Doch kann epistemischer Zugang zum Status der Erfüllung eines Desideratums ein wichtiger Schritt sein, um Fehler, Verzerrungen oder sonstige Probleme zu identifizieren und zu beheben. Verstehen fungiert dadurch, wie Markus Langer et al. es ausdrücken, als *Mediator* zwischen Erklärbarkeit und der Erfüllung gesellschaftlicher Desiderata (Langer et al., 2021).

Diese intuitiv plausible Annahme ist in der Erklärbarkeitsliteratur, zumindest implizit, weit verbreitet. Verstehen soll – unter anderem – (angemessenes) Vertrauen in KI-Systeme ermöglichen und Rechtssicherheit schaffen; es soll dabei helfen, die Verantwortlichen für das Scheitern eines Systems zu identifizieren und zur Rechenschaft zu ziehen; mithilfe von Verstehen sollen problematische Verzerrungen oder Fehler im System aufgedeckt und Entwickler:innen dabei unterstützt werden, diese zu korrigieren (Krishnan, 2020; Langer et al., 2021). Was genau es mit diesem Verstehen auf sich hat, und wie es die Erfüllung solch unterschiedlicher Desiderata ermöglichen kann, bleibt aber zumeist offen. Was muss

beispielsweise eine Systementwicklerin zu welchem Grad verstehen, um beurteilen zu können, ob ein KI-System in relevanten Hinsichten fair ist? Wie unterscheidet sich dieses Verständnis von jenem einer Endnutzerin ohne informatische Expertise, die ebenfalls ein Interesse daran hat, dass das System hinreichend fair ist? Und macht es einen Unterschied, ob ein KI-System bloß Zebrabeine zählen oder vielmehr Hautkrebs erkennen soll – also ob im fraglichen Kontext viel oder wenig auf dem Spiel steht?

Ich denke, dass wir eine überzeugende Antwort auf solche Fragen brauchen. Wenn Erklärbarkeit systembezogene Aspekte für unterschiedliche Stakeholder in verschiedenen (auch risikobehafteten) Kontexten verständlich machen und dadurch zur Erfüllung diverser Desiderata beitragen soll, braucht es mehr als vage Verweise auf ‚Verstehen‘ (Krishnan, 2020; Páez, 2019; Longo et al., 2024, Kap. 3.3.3). Es muss klar sein, worin dieser epistemische Zielzustand besteht, um Erklärbarkeitsmethoden entwickeln zu können, die tatsächlich ein effektives Mittel zur Erfüllung all dieser legitimen Anforderungen, Wünsche und Bedürfnisse darstellen, die in KI-Kontexten zum Tragen kommen. Zwar sind in den letzten Jahren einige Arbeiten erschienen, die philosophische Überlegungen zu Verstehen auf den XAI-Kontext übertragen (etwa Fleisher, 2022; Páez, 2019; Schmidt, 2026; Sullivan, 2022). Doch fehlt bislang ein philosophischer Beitrag, der die Mediatorrolle von Verstehen in XAI-Kontexten in den Mittelpunkt rückt und einen auf diese besondere Rolle zugeschnittenen Verstehensbegriff entwickelt.

1.1 Forschungsfragen

Diese Forschungslücke, die zwischen dem Black-Box-Problem, den Lösungsversuchen der XAI-Forschung und der Mediatorrolle von Verstehen verortet ist, motiviert mein Forschungsvorhaben. Konkret gehe ich in dieser Dissertation zwei Forschungsfragen nach. Die erste betrifft das Verstehen, das zwischen Erklärbarkeit und gesellschaftlichen Desiderata mediiert:

- (a) Welche Form des Verstehens ermöglicht Stakeholdern die Erfüllung gesellschaftlicher Desiderata in XAI-Kontexten?

Für die Beantwortung von Forschungsfrage (a) stütze ich mich nicht allein auf die klassische philosophische Methode der Begriffsanalyse, sondern verfolge vornehmlich einen begriffsgestaltenden Ansatz (meine Übersetzung von ‚*conceptual engineering*‘). Mithilfe einer an Rudolf Carnaps (1971) Methode angelehnten Explikation gestalte ich einen Verstehensbegriff, welcher der geschilderten Mediatorrolle Rechnung trägt und erklärt, inwiefern Verstehen als Mittel zur Erfüllung relevanter Desiderata fungieren kann.

Mein Ergebnis ist eine kontextsensitive Definition des Verstehens von KI-Systemen bzw. von zugehörigen systembezogenen Aspekten. Das entscheidende Merkmal dieses Verstehensbegriffs ist sein zweckgerichteter Charakter: Die Form des Verstehens, die mich

interessiert, ist kein Selbstzweck, sondern dient der Erfüllung legitimer gesellschaftlicher Desiderata. Ich nenne diese spezielle Verstehensform daher ‚instrumentelles Verstehen‘.

Meine Definition instrumentellen Verstehens besteht aus zwei notwendigen und zusammen hinreichenden Bedingungen. Die erste Bedingung verlangt vom instrumentell verstehenden Stakeholder ein gerechtfertigtes, moderat faktives Überzeugungssystem über systembezogene Aspekte. Die zweite Bedingung erfordert das Verfügen über verstehensbezogene Fähigkeiten. Ich identifiziere zudem drei verschiedene Kontextfaktoren, die sich auf die Bedingungen instrumentellen Verstehens auswirken – das Desideratum und die Rolle des Stakeholders sowie die kontextuellen Chancen bzw. Risiken.

Meine zweite Forschungsfrage nimmt den *Erwerb* instrumentellen Verstehens in den Blick. Ein wichtiges Mittel für den Verstehenserwerb in XAI-Kontexten sind die Ausgaben der oben erwähnten Erklärbarkeitsmethoden. Plausiblerweise kann instrumentelles Verstehen aber auch auf anderen Wegen erworben oder verbessert werden – etwa durch Erläuterungen von KI-Expert:innen, technische Systemdokumentationen oder Hintergrundinformationen zur Anwendungsdomäne des Systems. Ich fasse die unterschiedlichen Mittel, die für den Erwerb instrumentellen Verstehens genutzt werden können, unter der Bezeichnung ‚Verstehensermöglicher‘ zusammen.⁵ Doch wodurch zeichnet sich ein Verstehensermöglicher aus, der den Erwerb instrumentellen Verstehens begünstigt? Darum geht es in meiner zweiten Forschungsfrage:

- (b) Welche Eigenschaften muss ein Verstehensermöglicher aufweisen, um der Erlangung instrumentellen Verstehens zuträglich zu sein?

Zur Beantwortung von Forschungsfrage (b) ziehe ich Catherine Elgins (2017) Theorie der Exemplifikation heran. Mein Vorschlag lautet, dass ein epistemisch effektiver Verstehensermöglicher einem Stakeholder epistemischen Zugang zu relevanten systembezogenen Aspekten verschafft, indem er einige seiner Eigenschaften exemplifiziert und möglichst leicht zu interpretieren ist. Ich bezeichne solche epistemisch vorteilhaften Verstehensermöglicher als ‚EXEMPLARE‘ und charakterisiere sie anhand von fünf Eigenschaften: (1) EXEMPLARE heben exemplifizierte Merkmale hervor, (2) spielen inerte Merkmale herunter, (3) sie weisen einen klaren Geltungsbereich auf und exemplifizieren (4) kontextuell relevante sowie (5) für Menschen nachvollziehbare Merkmale.

Zusammengefasst: Forschungsfrage (a) betrifft die nähere Bestimmung jenes Verstehens, das der Mediatorfunktion zwischen XAI und Desiderata gerecht werden kann. Welchen epistemischen Zustand sollten Erklärbarkeitsmethoden und andere Verstehensermöglicher bei einem Stakeholder hervorbringen, damit dessen Desiderata erfüllt werden können? Ich nenne diesen epistemischen Zustand ‚instrumentelles Verstehen‘. Forschungsfrage (b) setzt sich hingegen mit den Mitteln auseinander, die für den Erwerb dieses instrumentellen Verstehens genutzt werden können. Ich bezeichne diese Mittel als ‚Verstehensermöglicher‘. Wie müssen Verstehensermöglicher beschaffen sein, damit ein Stakeholder instrumentelles

⁵Ich verdanke diese Bezeichnung Eva Schmidt.

Verstehen bestmöglich erwerben kann? Diese Dissertation entwickelt Antworten auf beide Fragen.

Die bisherigen Überlegungen lassen sich in Form eines Analyseschemas für XAI-Kontexte ausdrücken, das in Abbildung 1.1 dargestellt ist: Ein Stakeholder überwindet mithilfe eines oder mehrerer Verstehensermöglicher epistemische Opazität, die von unterschiedlichen Quellen rühren kann. Dadurch erwirbt er instrumentelles Verstehen von systembezogenen Aspekten. Dieses instrumentelle Verstehen ermöglicht dann die Erfüllung der gesellschaftlichen Desiderata dieses Stakeholders.

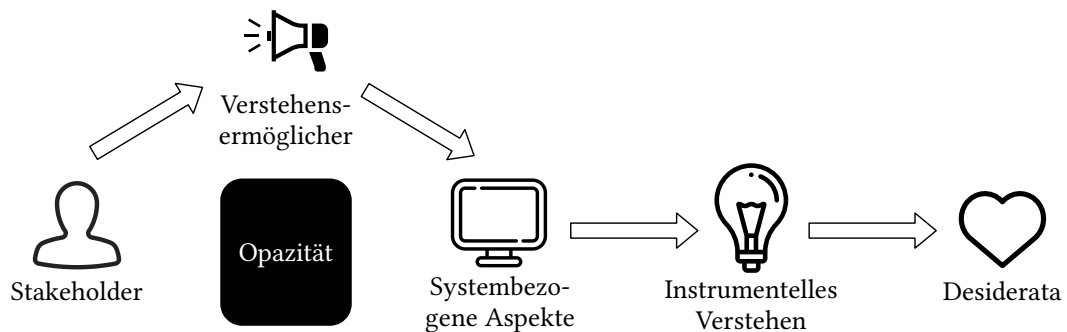


Abbildung 1.1: Übersicht über das hier zugrunde gelegte Analyseschema für XAI-Kontexte. Auf die verschiedenen Stakeholder in XAI-Kontexten geht Abschnitt 4.2.1 näher ein. Um Verstehensermöglicher geht es einführend in Abschnitt 3.1 und 3.3 sowie ausführlich in Kapitel 12. Verschiedene Quellen von epistemischer Opazität diskutiere ich in Abschnitt 2.3. Um systembezogene Aspekte als Verstehensobjekte geht es in Abschnitt 6.3. Instrumentellem Verstehen widmet sich der gesamte Teil II dieser Dissertation, mein Definitionsvorschlag findet sich in Abschnitt 11.1.1. Einen Überblick über die gesellschaftlichen Desiderata, die mein Vorhaben motivieren, bietet Abschnitt 4.2.3. Bildquelle: Eigene Darstellung unter Verwendung von Public-Domain-Grafiken. Siehe Teil IV für Details.

1.2 Aufbau der Arbeit

Meine Dissertation ist in drei Teile gegliedert. Aufgrund des interdisziplinären Charakters dieser Arbeit führt **Teil I: Grundlagen** zunächst in die Themengebiete Künstliche Intelligenz und Erklärbarkeit von KI-Systemen ein. In *Kapitel 2* gebe ich eine Übersicht über zentrale Methoden und Begriffe in den Bereichen KI und Maschinelles Lernen und stelle einige Quellen von epistemischer Opazität vor, welche die Verständlichkeit von KI-Systemen hemmen können. Anschließend legt *Kapitel 3* begriffliche Grundlagen zu Erklärbarkeit, stellt einige zentrale Kategorien informatischer Erklärbarkeitsmethoden vor und schlägt die Brücke zu philosophischen Überlegungen im Bereich Erklärung und Verstehen.

Anschließend folgt der ausführlichere **Teil II: Instrumentelles Verstehen**, der sich mit der Beantwortung von Forschungsfrage (a) befasst. *Kapitel 4* charakterisiert und motiviert die Carnap'sche Methode der Begriffsgestaltung, auf die ich mich in dieser Arbeit stütze. Ich präzisiere außerdem die Explikationsaufgabe und formuliere Adäquatheitsbedingungen, die mein Begriff instrumentellen Verstehens erfüllen muss. Anschließend steige ich in die philosophische Debatte zu Verstehen ein. *Kapitel 5* widmet sich der Gradierbarkeit von Verstehen und argumentiert, dass instrumentelles Verstehen nicht gradierbar ist. In *Kapitel 6* systematisiere ich unterschiedliche Kategorien von Objekten des Verstehens und schlage vor, dass diese Kategorien keine distinkten Verstehensarten markieren, sondern näherungsweise die Komplexität des Verstehensobjekts abbilden. Mein Fazit lautet, dass die Objekte instrumentellen Verstehens stets komplex sind. Anschließend argumentiere ich in *Kapitel 7*, dass instrumentelles Verstehen Überzeugung erfordert. *Kapitel 8* plädiert dafür, dass instrumentelles Verstehen moderat faktiv ist, also mit bestimmten falschen oder nur approximativ wahren Überzeugungen kompatibel ist. In *Kapitel 9* schlage ich vor, dass instrumentelles Verstehen je nach Kontext unterschiedliche Fähigkeiten erfordert, die sich insgesamt sechs Fähigkeitsclustern zuordnen lassen. *Kapitel 10* argumentiert, dass instrumentelles Verstehen Rechtfertigung erfordert und dass diese Rechtfertigung internalistisch und hybrid ist, also kohärentistische und fundamentalistische Elemente vereint. Aufbauend auf dieser ausführlichen Auseinandersetzung mit erkenntnis- und wissenschaftstheoretischen Debatten zu Verstehen schließt *Kapitel 11* meine Explikation ab. Ich führe das Explikat in Form meiner Definition instrumentellen Verstehens ein und evaluiere meinen Definitionsvorschlag anhand der zuvor aufgestellten Adäquatheitskriterien.

Schließlich ist **Teil III: Verstehensermöglicher** der Beantwortung von Forschungsfrage (b) gewidmet. In *Kapitel 12* charakterisiere ich EXEMPLARE als Verstehensermöglicher, die dem Erwerb instrumentellen Verstehens besonders zuträglich sind. Weiterhin wende ich meinen Vorschlag auf mehrere Erklärbarkeitsmethoden an, die genutzt werden, um Aspekte von Bildklassifikationssystemen verständlich zu machen. Anhand dieser Beispiele zeige ich auf, dass die Charakteristika von EXEMPLAREN genutzt werden können, um Verstehensermöglicher hinsichtlich ihrer Eignung für den Verstehenserwerb zu evaluieren. Im Anschluss fasst die Konklusion in *Kapitel 13* meine Ergebnisse zusammen und zeigt weiterführende Fragestellungen auf, die sich aus meiner Untersuchung ergeben haben.

Teil I

Grundlagen

2 Künstliche Intelligenz

Künstliche Intelligenz ist ein Sammelbegriff für ein interdisziplinäres Forschungsfeld, das sich mit der Erforschung und Erschaffung von intelligenten Artefakten, etwa in Form von Computerprogrammen oder Robotern, befasst.⁶ Innerhalb der KI-Forschung lassen sich zwei Hauptrichtungen unterscheiden (Görz et al., 2021, S. 2): Auf der einen Seite steht der psychologisch-kognitionswissenschaftliche Zweig, der versucht, menschliche Kognition und Intelligenz durch Artefakte zu modellieren und zu erforschen. Auf der anderen Seite steht der ingenieurwissenschaftliche Zweig, der Methoden entwickelt, um „bestimmte menschliche Wahrnehmungs- und Verstandesleistungen maschinell verfügbar und praktisch nutzbar [zu] machen“ (ebd.) – unabhängig davon, ob diese Methoden der menschlichen Kognition ähneln.

Im Kontext dieser Arbeit beschränke ich mich auf die ingenieurwissenschaftliche Perspektive. KI-Systeme sind demnach Werkzeuge, die Menschen dabei helfen sollen, spezifische Aufgaben zu lösen.⁷ Dazu passt die folgende Definition aus der ‚Verordnung über künstliche Intelligenz‘ der Europäischen Union:

„KI-System“ [bezeichnet] 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[.] (Artikel 3 Absatz (1) Verordnung (EU) 2024/1689)

Die Verordnung bestimmt KI-Systeme also als technologische Artefakte, die mit unterschiedlichem Autonomiegrad Eingaben in Ausgaben umwandeln. Wie genau diese Umwandlung stattfindet, hängt von der spezifischen Systemarchitektur ab. Die wichtigsten Systemarchitekturen lassen sich anhand zweier Unterscheidungen kategorisieren: (1) symbolische Ansätze versus subsymbolische Ansätze sowie (2) konventionelle KI versus Maschinelles Lernen. Das Verhältnis der vier daraus resultierenden Kategorien ist in Abbildung 2.1 dargestellt.

Die erste Unterscheidung, also jene zwischen symbolischer und subsymbolischer KI, betrifft die Art, wie Information in einem KI-System repräsentiert wird. *Symbolische KI-*

⁶Vielen Dank an Timo Speith, der eine frühere Version dieses Kapitels auf informatische Korrektheit geprüft hat. Etwaige verbleibende Fehler sind meine eigenen.

⁷Diese Perspektive betont auch, dass KI nicht wie eine unaufhaltsame ‚Naturgewalt‘ ist, sondern dass letztlich Menschen darüber entscheiden, wie diese Technologie gestaltet, verwendet und reguliert wird.

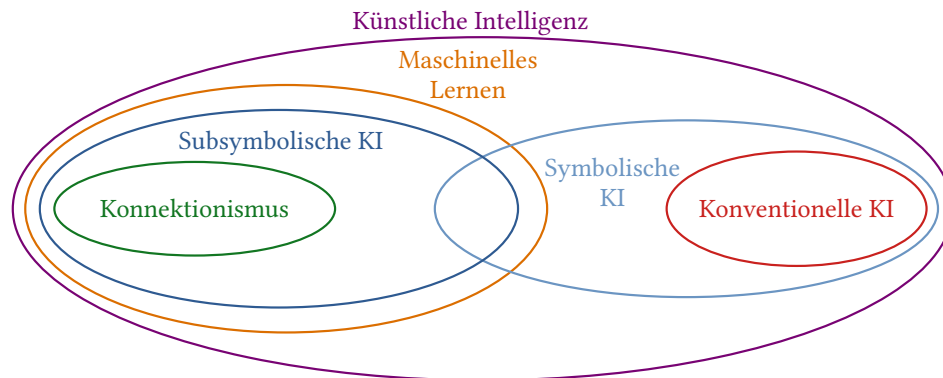


Abbildung 2.1: Vereinfachte Darstellung des Verhältnisses zwischen einigen prominenten Forschungsbereichen innerhalb der informatischen KI-Forschung. Die für diese Arbeit maßgeblichen Unterscheidungen verlaufen zwischen (1) symbolischer und subsymbolischer KI sowie zwischen (2) konventioneller KI und Maschinellern Lernen. Bildquelle: Eigene Darstellung. Siehe Teil IV für Details.

Systeme manipulieren (in der Regel menschenverständliche) Symbole. Typische Beispiele sind Expertensysteme oder Entscheidungsbäume. *Subsymbolische* Systeme arbeiten dagegen mit distribuierten, numerischen Repräsentationen, die für Menschen in der Regel nicht unmittelbar verständlich sind. Dazu zählen etwa auf Fisher-Vektoren basierende Systeme oder Stützvektormaschinen. Die bekanntesten Vertreter sind aber Systeme, die dem Konnektionismus zuzurechnen sind, also künstliche neuronale Netze. Darüber hinaus existieren auch kombinierte Ansätze, die symbolische und subsymbolische Repräsentationen miteinander verbinden.

Die zweite Unterscheidung verläuft zwischen konventioneller KI und Maschinellern Lernen (ML). Sie betrifft die Art, wie das KI-System entwickelt wird. *Konventionelle* KI-Systeme befolgen einzig explizit einprogrammierte Abläufe. Ein bekannter Vertreter sind die oben genannten Expertensysteme. KI-Systeme hingegen, die auf Methoden des *Maschinellen Lernens* beruhen, entwickeln selbst eine Lösung für eine bestimmte Aufgabe. Solche Systeme sind maßgeblich für die jüngsten Erfolge der KI-Forschung verantwortlich. Ein wichtiger Ansatz des Maschinellen Lernens ist das ‚Deep Learning‘ mit tiefen künstlichen neuronalen Netzen.

Die vier Kategorien, die sich aus diesen beiden Unterscheidungen ergeben, überlappen einander (siehe Abbildung 2.1). Da konventionelle KI händisch programmiert wird, beruht sie auf symbolischen Repräsentationen. Maschinelles Lernen kann dagegen sowohl mit manchen symbolischen (etwa Entscheidungsbäumen) als auch subsymbolischen Systemen (wie neuronalen Netzen) durchgeführt werden.

Die ersten beiden Abschnitte dieses Kapitels gehen näher auf diese Kategorien ein: In Abschnitt 2.1 erläutere ich die Unterscheidung zwischen symbolischen und subsymbolischen

Systemen. Abschnitt 2.2 betrachtet die dazu quer verlaufenden Kategorien der konventionellen KI einerseits und des Maschinellen Lernens andererseits. Diese informatischen Hintergründe bilden die Grundlage, um in Abschnitt 2.3 verschiedene Ursachen für die epistemische Opazität von KI-Systemen und damit die Notwendigkeit von erklärbarer KI zu verdeutlichen.

2.1 Symbolische versus subsymbolische Ansätze

Die Unterscheidung zwischen dem symbolischen und subsymbolischen Paradigma geht maßgeblich auf Paul Smolensky zurück (Smolensky, 1988). Gemäß dem symbolischen Paradigma werden kognitive Prozesse (bzw. Prozesse, die in einem KI-System ablaufen) durch Symbole modelliert. Laut Smolensky sind solche Symbole semantisch in dem Sinne, dass sie auf externe Objekte verweisen, und syntaktisch in dem Sinne, dass sie durch Symbolmanipulation verarbeitet werden (ebd., S. 3). Damit ähnelt die Funktionsweise eines symbolischen KI-Systems einer Formalisierung der bewussten Denkvorgänge von Menschen und ist entsprechend intuitiv.

Systeme hingegen, die gemäß dem subsymbolischen Paradigma aufgebaut sind, modellieren kognitive Prozesse durch die Konstituenten der obigen Symbole. Dies hat Auswirkungen auf die Semantik von Subsymbolen: Eine Entität, die im symbolischen Paradigma durch ein einzelnes Symbol repräsentiert wird, wird im subsymbolischen Paradigma durch eine größere Anzahl von Subsymbolen repräsentiert (ebd.). Dies führt zu einer distribuierten Repräsentation, auf die ich unten näher eingehen werde. Subsymbole haben auch eine abweichende Syntax: Anstatt durch Symbolmanipulation werden Subsymbole durch numerische Operationen verarbeitet. Smolensky zufolge lassen sich Subsymbole, betrachtet auf einem höheren Abstraktionslevel, wiederum zu Symbolen zusammensetzen (ebd.).⁸

Die Unterscheidung zwischen symbolischen und subsymbolischen KI-Systemen lässt sich am besten anhand von Beispielen verdeutlichen. Ich gehe nachfolgend näher auf einen einfachen Entscheidungsbaum als Beispiel für ein symbolisches System (Abschnitt 2.1.1) sowie auf neuronale Netze als Beispiel für subsymbolische KI ein (Abschnitt 2.1.2).

2.1.1 Symbolische KI

Ein Entscheidungsbaum prüft systematisch, ob bestimmte Attribute auf eine Eingabe zutreffen (Beierle & Kern-Isberner, 2019, S. 107). Ein simples Beispiel ist die Auswahl eines geeigneten Fortbewegungsmittels (siehe Abbildung 2.2). Der Entscheidungsknoten am Stamm fragt die Distanz zwischen Start- und Zielpunkt ab. Beträgt diese weniger als fünf

⁸Eine entsprechende Hoffnung spiegelt sich heute in Erklärbarkeitsansätzen wider, die versuchen, die Repräsentationen tiefer neuronaler Netze mit menschlichen Begriffen zusammenzubringen (z. B. Achibat et al., 2023; Bau et al., 2020; Fong & Vedaldi, 2018; B. Kim et al., 2018; eine kritische Übersicht bietet Freiesleben, 2026).

Kilometer, führt der nächste Zweig zu einem weiteren Entscheidungsknoten, der nach dem Vorhandensein von Zeitdruck fragt, und so weiter.

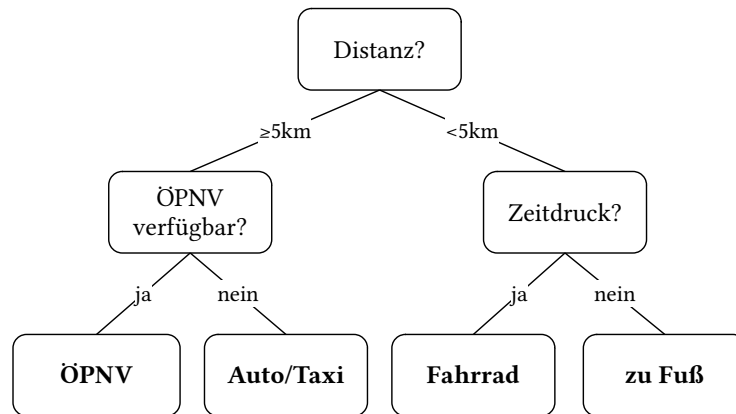


Abbildung 2.2: Ein Beispiel für einen einfachen Entscheidungsbaum, mit dem sich abschätzen lässt, welches Fortbewegungsmittel man in Abhängigkeit von der zu überbrückenden Distanz, dem Vorhandensein von Zeitdruck und der Verfügbarkeit öffentlicher Verkehrsmittel nutzen soll. Bildquelle: Eigene Darstellung. Siehe Teil IV für Details.

Entscheidungsbäume erlauben es, jede Ausgabe des Systems unmittelbar nachzuvollziehen: Sie ergibt sich aus dem Pfad, den das Eingabedatum in der Baumstruktur genommen hat. Unter der Voraussetzung, dass die Daten, auf deren Grundlage der Entscheidungsbaum generiert wurde, symbolisch strukturiert sind, gilt dies auch für die Attribute, die der Entscheidungsbaum abprüft. Der Pfad, der hinter einer bestimmten Ausgabe steht, besteht dann aus einer Aneinanderreihung menschenverständlicher Symbole (z. B. ‚Distanz‘ oder ‚Zeitdruck‘). Daher werden Entscheidungsbaume häufig als inhärent interpretierbar angesehen (siehe Abschnitt 3.2.1).

2.1.2 Subsymbolische KI

Subsymbolische Repräsentationen werden unter anderem von konnektionistischen Modellen genutzt, also künstlichen neuronalen Netzwerken. Die Architektur dieser Netzwerke ist lose vom Aufbau des Gehirns inspiriert und viele zugehörige Ausdrücke sind biologischer Terminologie entlehnt. Passend dazu werden künstliche neuronale Netze meist als tatsächliche Netzwerke dargestellt, durch die ein Signal wandert (siehe Abbildung 2.3a). Dabei handelt es sich jedoch lediglich um eine nützliche Metapher – eigentlich sind diese Modelle „weder neural, noch Netzwerke“ (Chollet, 2021, S. 434, meine Übersetzung),

sondern beschreiben eine Abfolge mathematischer Operationen.⁹ Entsprechend sind auch subsymbolische Repräsentationen rein numerisch. Um zu erläutern, was mit einer solchen Repräsentation gemeint ist, bedarf es zunächst einer Darstellung der Funktionsweise von künstlichen neuronalen Netzen.

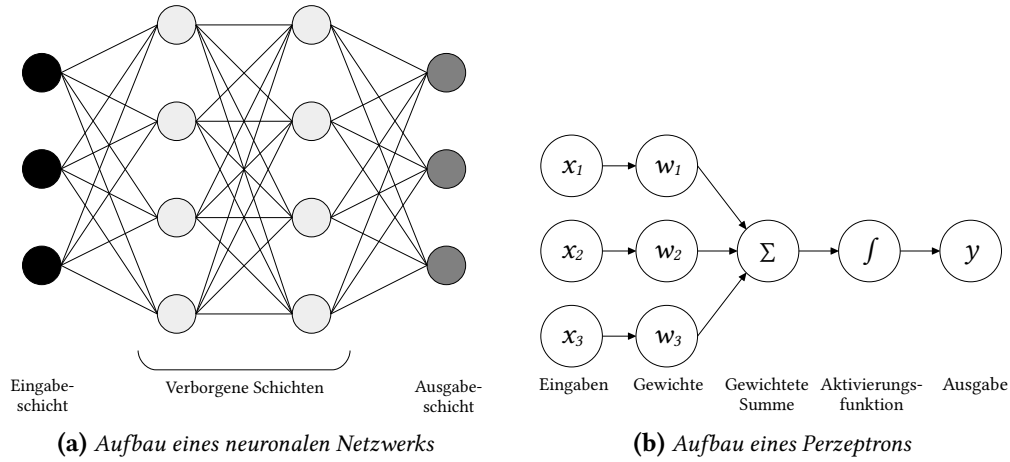


Abbildung 2.3: Visualisierung der Grundstruktur künstlicher neuronaler Netzwerke. **(a)** Der Aufbau eines einfachen neuronalen Netzwerks bestehend aus einer Eingabeschicht, mehreren verborgenen Schichten sowie einer Ausgabeschicht. Die Kreise stehen für Neuronen, die Linien für Kanten. Bildquelle: Eigene Darstellung, adaptiert von einer Public-Domain-Grafik. Siehe Teil IV für Details. **(b)** Darstellung der Signalverarbeitung innerhalb eines Perzeptrons, dem simpleren Vorgänger heutiger neuronaler Netze. Die Eingaben des Neurons werden gewichtet und aufsummiert. Durch eine Aktivierungsfunktion wird diese gewichtete Summe transformiert und an das nächste Neuron weitergegeben. Bildquelle: Eigene Darstellung, adaptiert von Maier et al. (2019). Siehe Teil IV für Details.

Die wichtigsten Elemente eines neuronalen Netzwerks sind die Neuronen sowie deren Verknüpfungen untereinander, die so genannten Kanten. Die Kanten beschreiben die Pfade, entlang derer eine numerische Eingabe – das Signal – zwischen Neuronen weitergegeben

⁹Neuronale Netze sind eine mathematische Methode – nicht mehr und nicht weniger. Eine akkurate Darstellung eines künstlichen neuronalen Netzes sieht daher eher so aus (Angelov et al., 2021, S. 3):

$$y_p = f_n \left(\sum_q \left(w_{pq} f_{n-1} \left(\dots f_2 \left(\sum_i \left(w_{ri} f_1 \left(\sum_j (w_{ij} x_j) \right) \right) \right) \right) \right) \right)$$

y_p ist eine einzelne Ausgabe, w sind die Kantengewichte, j ist die Anzahl der Eingabefeatures, i ist die Anzahl der Neuronen in einer verborgenen Schicht, f_1 ist die Aktivierungsfunktion der Eingabeschicht, f_2 steht für die Aktivierungsfunktion einer verborgenen und f_n für die Aktivierungsfunktion der Ausgabeschicht. Ein künstliches neuronales Netz ähnelt aufgrund der Verschachtelung der Rechenschritte einer Matroschkapuppe (ebd.).

wird. Die Neuronen sind in Subpopulationen angeordnet, den Schichten. Innerhalb einer solchen Schicht sind die Neuronen nicht untereinander verknüpft, sondern nur mit allen oder einigen Neuronen der vor- sowie nachgelagerten Schichten. Daraus ergibt sich eine spezifische Netzwerkstruktur (siehe Abbildung 2.3a): Typischerweise besteht ein künstliches neuronales Netz aus einer Eingabeschicht, die den Input entgegennimmt; mehreren ‚verborgenen Schichten‘, in denen das Eingangssignal verarbeitet wird; und schließlich einer Ausgabeschicht, die das Gesamtergebnis ausgibt, das sich aus den Einzelberechnungen der Neuronen des Netzes ergeben hat (Goodfellow et al., 2016, Kap. 6).

Zur Illustration genügt eine vereinfachte Beschreibung eines vorwärts gekoppelten Netzwerks (engl. *feedforward network*), in dem das Signal lediglich in eine Richtung – von der Eingabe- zur Ausgabeschicht – weitergeleitet wird. Bei einigen der wichtigsten Netzwerkarchitekturen handelt es sich (partiell) um vorwärtsgekoppelte Netzwerke – etwa bei CNN (*convolutional neural networks*), die sich besonders gut für die Bildklassifikation eignen, oder bei Bestandteilen von Transformer-Netzwerken, die in großen Sprachmodellen zum Einsatz kommen. In einem vorwärtsgekoppelten Netzwerk steht jedes Neuron und jede Kante für eine bestimmte mathematische Operation, die mit dem Signal durchgeführt wird, um sukzessive relevante von irrelevanten Informationen zu trennen (siehe Abbildung 2.3b). Wenn das Signal, bildlich gesprochen, eine Kante passiert, wird es mit dem Kantengewicht multipliziert und dadurch entweder verstärkt oder abgeschwächt. Dieser gewichtete Wert erreicht dann ein Neuron. Dort laufen mehrere Kanten zusammen. Ein Neuron summiert alle ankommenden gewichteten Werte auf. Diese gewichtete Summe wird dann in eine Aktivierungsfunktion eingespeist. Eine verbreitete Aktivierungsfunktion ist beispielsweise die Sigmoidfunktion, die jede Eingabe in einen Wert zwischen Null und Eins umwandelt. Der durch die Aktivierungsfunktion transformierte Wert ist dann die Ausgabe des Neurons, oder seine Aktivierung. Dieser Aktivierungswert wird über eine gewichtete Kante an das nachgelagerte Neuron weitergeleitet. Nun wiederholt sich der beschriebene Ablauf, bis die Ausgabeschicht erreicht ist.

Auf dieser Grundlage lässt sich beschreiben, was eine subsymbolische Repräsentation ist: Es handelt sich um ein spezifisches, numerisches Aktivierungsmuster eines neuronalen Netzwerks. Dieses Muster besteht aus den Aktivierungsgraden der Neuronen, wenn diese eine bestimmte Eingabe verarbeiten. Das Aktivierungsmuster kann sich prinzipiell auf ein einzelnes Neuron beschränken, ist häufig aber auf mehrere oder gar alle Neuronen eines Netzwerks distribuiert (Sun, 2014, S. 112–113). Das erlaubt eine effizientere Repräsentation: Anstatt beispielsweise jeweils ein Neuron für die Repräsentation von schwarzen Dreiecken, weißen Dreiecken, schwarzen Quadraten und weißen Quadraten zu verwenden, genügt je ein Neuron für ‚schwarz‘, ‚weiß‘, ‚Dreieck‘ und ‚Quadrat‘. Diese Neuronen können dann auch genutzt werden, um andere schwarze bzw. weiße Objekte oder andersfarbige Quadrate bzw. Dreiecke zu erkennen. Außerdem vervielfältigt sich dadurch die Zahl der Trainingsdaten, die für das Erlernen solcher Repräsentationen zur Verfügung stehen: Das Neuron für ‚schwarz‘ kann auch mit Beispielen schwarzer Kreise, Fahrräder oder Bücher trainiert werden (Goodfellow et al., 2016, S. 16–17).

Diese Form der Repräsentation macht tiefe neuronale Netze zu mächtigen Werkzeugen für die Lösung anspruchsvoller Aufgaben aus dem Bereich Computervision oder Computerlinguistik (ebd., Kap. 1.2.4). Derart komplexe Funktionen lassen sich nicht sinnvoll händisch programmieren. Stattdessen sind subsymbolische Repräsentationen das Ergebnis von Verfahren des Maschinellen Lernens.

2.2 Konventionelle KI versus Maschinelles Lernen

ML-gestützte Verfahren haben konventionelle KI-Systeme mittlerweile in vielen Bereichen überflügelt. Dabei haben sich die Lernalgorithmen seit den 1980er Jahren nur geringfügig verändert. Neben der Verbesserung bestehender Algorithmen waren zwei Entwicklungen entscheidend für die Fortschritte im Bereich ML: die Verfügbarkeit großer, für das Training erforderlicher Datensätze („Big Data“) seit Beginn des 21. Jahrhunderts sowie die gewachsenen computationalen Ressourcen, welche die Verwendung deutlich größerer Modelle ermöglichen (Chollet, 2021, S. 20–22; Goodfellow et al., 2016, S. 18–22). Konventionelle KI-Systeme sind deswegen aber nicht obsolet. Beide Ansätze haben ihre Vor- und Nachteile und sind entsprechend für verschiedene Anwendungsbereiche geeignet (Boden, 2014, Kap. 4.3; S. Russell & Norvig, 2021, S. 1018).

ML-basierte Systeme eignen sich besonders gut für die automatische Mustererkennung in großen Datensätzen (z. B. Data-Mining) sowie für die Lösung unzureichend verstandener Probleme, die keine konventionelle Programmierung erlauben (T. Mitchell, 1997, S. 17). Aufgrund dieser Stärken werden ML-gestützte Systeme in immer neuen Anwendungsbereichen erfolgreich eingesetzt. Beispielsweise fallen in medizinischen Kontexten immense Mengen unterschiedlicher Daten an, die für das Training von ML-Systemen verwendet werden können, etwa zur Erkennung von Hautkrebs auf Grundlage von Bilddaten (Esteva et al., 2017). Maschinelles Lernen steckt auch hinter vielen der jüngeren Fortschritte im Bereich des autonomen Fahrens. ML-basierte KI-Systeme analysieren in Echtzeit die Sensordaten des Fahrzeugs, etwa zur Positionsbestimmung oder um Objekte bzw. Personen zu erkennen (Grigorescu et al., 2020). Diverse andere Anwendungsfälle profitieren von der ML-basierten Auswertung großer Datenmengen, z. B. die Entdeckung von Exoplaneten auf Grundlage von Weltraumteleskopdaten (Shallue & Vanderburg, 2018) oder die Vorhersage von Proteinstrukturen (Abramson et al., 2024).

Konventionelle KI-Systeme können dagegen besonders gut hierarchische und sequenzielle Ordnungen abbilden (Boden, 2014, S. 93). Aufgrund des vergleichsweise hohen Entwicklungsaufwands eignen sie sich vor allem für klar abgegrenzte, statische Anwendungsbereiche, in denen sich wiederholende Aufgaben anhand fester Regeln gelöst werden können. Dort erweisen sich konventionelle KI-Technologien bis heute als effektive Werkzeuge (Beierle & Kern-Isberner, 2019, S. 14; Buchanan et al., 2018, S. 96). Zum Beispiel existieren Expertensysteme für zahlreiche medizinische Anwendungen, diese basieren entweder allein auf konventioneller KI oder kombinieren diese mit Maschinellern Lernen

(Saibene et al., 2021).

Ich gehe in Abschnitt 2.2.1 zunächst näher auf ein klassisches Beispiel konventioneller KI ein, die so genannten wissensbasierten Systeme. Anschließend erläutere ich in Abschnitt 2.2.2 zwei exemplarische Verfahren des Maschinellen Lernens.

2.2.1 Konventionelle KI

Eine wichtige Systemfamilie innerhalb der konventionellen KI sind die so genannten wissensbasierten Systeme.¹⁰ Wissensbasierte Systeme zeichnen sich durch zwei Komponenten aus: Zum einen die *Wissensbasis*, in der anwendungsspezifisches Wissen gespeichert ist; zum anderen die *Wissensverarbeitung*, welche anwendungsunspezifische Problemlösestrategien enthält (Beierle & Kern-Isberner, 2019, S. 11). Typischerweise wird das Wissen in einer Wissensbasis in Form von Wenn-dann-Regeln dargestellt (siehe dazu ebd., Kap. 4). Diese Regeln können zum Beispiel aus durch ML erzeugten Entscheidungsbäumen (siehe Abschnitt 2.1) generiert (ebd., S. 109) oder auf Grundlage von Expertenwissen formuliert werden.

Letztere Strategie ist typisch für einen bekannten Vertreter wissensbasierter Systeme, die so genannten Expertensysteme. Ein Expertensystem soll Expertenwissen zu einem bestimmten Problembereich repräsentieren und (z. B. für Nichtexpert:innen) nutzbar machen. Margaret Boden gibt folgendes Beispiel für eine Wenn-dann-Regel aus einem medizinischen Expertensystem:

IF *these* symptoms are observed in the patient, AND the patient is an adult male, THEN infer that *this* bacterium is responsible, AND recommend *this* drug, given in *that* dosage for *that* length of time, as the treatment. (Boden, 2014, S. 91; Hervorhebung im Original)

Natürlich muss eine solche aus dem Fachwissen menschlicher Expert:innen generierte Regel in eine maschinenlesbare, formale Sprache übersetzt werden, damit sie von Expertensystemen verarbeitet werden kann (ebd., S. 89–90; Beierle & Kern-Isberner, 2019, S. 73–77). Doch auch dann korrespondieren die Elemente der Regeln menschenverständlichen Begriffen (Buchanan et al., 2018, S. 84). Eine Rückübersetzung in natürliche Sprache ist daher gut machbar, das nötige (informatische) Hintergrundwissen vorausgesetzt.¹¹

Auf dieser Grundlage können auch recht einfach Erklärungen für Entscheidungen generiert werden (ebd.). Expertensysteme enthalten normalerweise eine so genannte Erklärungskomponente (Beierle & Kern-Isberner, 2019, S. 18, 88), welche die herangezogenen Regeln benennt und dadurch ein Argument für das Ergebnis produziert. In gleicher Weise kann auch das gesamte System inspiziert werden (Buchanan et al., 2018, S. 94–95). Das

¹⁰Da es sich um einen informatischen Begriff handelt, darf *Wissen* in diesem Kontext nicht erkenntnistheoretisch interpretiert werden.

¹¹Ein reales Beispiel für eine solche Regel samt Übersetzung in natürliche Sprache findet sich in der Publikation zum damals äußerst einflussreichen medizinischen Expertensystem MYCIN (Shortliffe et al., 1973, S. 547).

Vorhandensein einer Erklärungskomponente reicht zwar nicht immer aus, um Verständlichkeit sicherzustellen. Doch wird durch die Verwendung symbolischer Repräsentationen eine wichtige Quelle von Opazität umschifft, die typisch für subsymbolische Systeme ist (mehr dazu in Abschnitt 2.3.1).

Expertensysteme erlebten in den 1980er Jahren einen regelrechten Boom, konnten die hohen Erwartungen letztlich aber nicht erfüllen (S. Russell & Norvig, 2021, Kap. 1.3.4). Vor diesem Hintergrund werden konventionelle KI im Allgemeinen und Expertensysteme im Besonderen mitunter als gescheitert betrachtet. Wie oben dargestellt, ist diese Behauptung so pauschal nicht haltbar. Konventionelle KI-Systeme, einschließlich Expertensystemen, werden bis heute in verschiedensten Anwendungsbereichen eingesetzt (Buchanan et al., 2018, S. 95–98, bieten diverse Beispiele). Tatsächlich ist konventionelle KI heute derart selbstverständlich, dass sie mitunter nicht mehr als KI-Technologie erkannt wird (Boden, 2014, S. 101). Wie es einer der Gründungsväter der KI-Forschung, John McCarthy, einst ausgedrückt haben soll: „As soon as it works, no one calls it AI any more“ (zitiert in Meyer, 2011, o. S.). Nichtsdestotrotz haben sich ML-Verfahren in vielen Anwendungsbereichen gegenüber konventionellen Systemen durchgesetzt.

2.2.2 Maschinelles Lernen

Maschinelles Lernen umfasst Methoden, die es Maschinen erlauben, auf Grundlage von Beispieldaten zu lernen. ‚Lernen‘ heißt in diesem Kontext, dass sich die Performance der Maschine bei der Lösung einer bestimmten Aufgabe verbessert (S. Russell & Norvig, 2021, S. 651). Die Performance kann je nach Aufgabe durch verschiedene Metriken gemessen werden, z. B. durch die mittlere quadratische Abweichung für Regressionsaufgaben oder die Genauigkeit für Klassifikationsaufgaben (ich gehe unten näher auf Klassifikation und Regression ein). Die Grundlage für diesen Lernprozess bildet die ‚Erfahrung‘ der Maschine. Damit sind die Daten gemeint, mit denen das System trainiert wird. Der entscheidende Unterschied zu konventionellen Programmierverfahren besteht darin, dass sich das System selbst verbessert. Anders als zum Beispiel bei Expertensystemen muss die Lösung für eine Aufgabe nicht durch Programmierer:innen vorgegeben werden. Stattdessen kann ein System durch ML selbst eine Lösung entwickeln, indem es während der Trainingsphase seine Leistungsfähigkeit in Hinblick auf die Aufgabe steigert. Eine klassische Definition dieses Verfahrens stammt von Tom Mitchell:

A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .¹² (T. Mitchell, 1997, S. 2; Hervorhebung im Original)

Der Grundprinzip Maschinellen Lernens ist die Mustererkennung, also das Auffinden von Korrelationen innerhalb der (‚Erfahrungs-‘)Daten. Anders als konventionell programmierte

¹²Siehe T. Mitchell (1997, Kap. 1.2) und Goodfellow et al. (2016, S. 96–104) für eine Diskussion der einzelnen Elemente dieser Definition.

Systeme basieren Verfahren des Maschinellen Lernens daher auf Wahrscheinlichkeiten anstatt auf logischen Regeln. Während des Lernprozesses werden automatisiert solche Korrelationsmuster identifiziert. Diese Muster werden dann herangezogen, um die Aufgabe zu lösen, d. h. probabilistische Aussagen über neue Daten zu treffen. Es existieren verschiedene Lernverfahren, die sich jeweils für andere Aufgaben und Daten eignen. Ich stelle im Folgenden zwei dieser Lernverfahren vor: überwachtes sowie unüberwachtes Lernen.¹³

Überwachtes Lernen

Beim überwachten Lernen bestehen die Trainingsdaten aus Eingabe-Ausgabe-Paaren, also aus Daten, die mit einem Label versehen sind. Auf Grundlage solcher Eingabe-Ausgabe-Paare erlernt der Algorithmus eine (symbolische oder subsymbolische) Regel, anhand derer er neuen Dateneingaben (die nicht in den Trainingsdaten enthalten waren) eine passende Ausgabe in Form eines Labels zuordnen kann.

Die beiden Hauptaufgaben im überwachten Lernen sind Regression und Klassifikation (S. Russell & Norvig, 2021, S. 652). Bei einer Klassifikationsaufgabe soll ein Computerprogramm einen Input der richtigen Kategorie aus einer vorgegebenen Menge von Kategorien zuordnen. Ein Standardbeispiel ist die Erkennung handgeschriebener Ziffern. Die Aufgabe des Algorithmus besteht darin, einer Eingabe (z. B. das Bild einer handgeschriebenen Drei) eine Ausgabe zuzuweisen (z. B. die Klasse ‚3‘). Es werden dabei keine Zwischenstufen zwischen den Bildklassen unterschieden. Die Zielvariable von Klassifikationsaufgaben ist also diskret/kategorial bzw. qualitativ. Bei einer Regressionsaufgabe soll das Computerprogramm einem Input einen numerischen Wert zuordnen. Zum Beispiel könnte ein Programm den Verkaufswert eines Grundstücks schätzen. Die Eingabe wären dann verschiedene Angaben zum Grundstück, die Ausgabe der Schätzpreis. Die Ausgabe könnte dann jeden Wert zwischen, sagen wir, 0€ und unendlich annehmen. Die Zielvariable von Regressionsaufgaben ist also kontinuierlich bzw. quantitativ.

Aus erkenntnistheoretischer Perspektive hat überwachtes Lernen den Vorteil, dass wenigstens die Ausgabe (z. B. ‚3‘ oder ‚250.000€‘) von Menschen vorgegeben und damit grundsätzlich interpretierbar ist. Es kann aber sein, dass die erlernte Regel, anhand derer neue Daten einer Ausgabe zugeordnet werden, nicht ohne weiteres verständlich ist – etwa,

¹³Ich lasse aus Platzgründen unter anderem ein Lernverfahren außer Acht, das aufgrund des Erfolgs großer Sprachmodelle in den letzten Jahren an Bedeutung gewonnen hat: das selbstüberwachte Lernen. Dabei werden während des ‚Pretext‘ so genannte ‚Pseudo-Label‘ aus den Trainingsdaten generiert. Diese Pseudo-Label sollen dann wie beim überwachten Lernen vorhergesagt werden (LeCun & Misra, 2021). Dadurch erlernen KI-Systeme bestenfalls robuste subsymbolische Repräsentationen, die sich zur Generation von Daten, wie Text oder Bildern, eignen. Ein weiteres wichtiges Lernverfahren, auf das ich nicht näher eingehen kann, ist Bestärkungslernen (engl. *reinforcement learning*). Dabei lernt das System, wie ein bestimmtes Ziel (eine ‚Belohnung‘) erreicht und ein nicht wünschenswerter Ausgang (eine ‚Strafe‘) vermieden werden kann (S. Russell & Norvig, 2021, Kap. 22). Dadurch können KI-Systeme z. B. lernen, Videospiele zu meistern oder Roboter zu steuern.

weil es sich um ein subsymbolisches Modell handelt, oder weil der (symbolische) Entscheidungspfad sehr kompliziert ist. An dieser Stelle setzen dann Methoden der erklärbaren KI an (siehe Kapitel 3).

Unüberwachtes Lernen

Beim unüberwachten Lernen enthalten die Trainingsdaten keine Zuordnung zu einer bestimmten Ausgabe, es ist also keine Zielvariable verfügbar (vgl. ebd., S. 738). Stattdessen analysiert der Algorithmus die Struktur der Daten und extrahiert nützliche Eigenschaften. Hier lassen sich abermals zwei prominente Aufgaben unterscheiden: Clusteranalyse und Assoziationsanalyse. Bei einer Clusteranalyse identifiziert der Algorithmus eigenständig Ähnlichkeiten zwischen den Datenpunkten des Datensatzes und teilt diese entsprechend in Gruppen ein. Beispielsweise lassen sich in astronomischen Daten über Sterne verschiedene Cluster identifizieren, die dann von Astronom:innen als ‚Weißer Zwerg‘ oder ‚Roter Riese‘ bezeichnet werden (ebd.). Bei der Assoziationsanalyse werden Assoziationsregeln zwischen den Datenpunkten formuliert. Ein klassisches Beispiel ist die Warenkorbanalyse. Als Eingabe dienen Daten zu getätigten Einkäufen von Kund:innen. Die Ausgabe sind Assoziationsregeln in Form von Zusammenhängen innerhalb dieser Daten, wie etwa ‚32% der Personen, die Backpulver kauften, kauften auch Eier‘. Die durch unüberwachtes Lernen erlernten Cluster bzw. Assoziationsregeln können dann genutzt werden, um neue Eingaben zu klassifizieren.

Unüberwachte Lernverfahren bergen eigene epistemische Herausforderungen. Eine Schwierigkeit besteht darin, dass es beim unüberwachten Lernen häufig keinen objektiven Gradmesser des Erfolgs gibt (von Luxburg et al., 2012). Bei überwachten Lernverfahren kann z. B. die Quote der korrekt klassifizierten Instanzen genutzt werden, um den Lernerfolg des Modells zu bewerten. Es ist jedoch viel schwieriger zu beurteilen, ob durch unüberwachtes Lernen aufgefundene Datencluster oder Assoziationsregeln valide sind. Häufig werden subjektiv-qualitative Merkmale, wie ‚Interessantheit‘ oder ‚Verständlichkeit‘, für die Evaluation des Modells herangezogen. Bei der Clusteranalyse muss sogar der Output selbst erst interpretiert werden. Der Algorithmus nummeriert die aufgefundenen Cluster lediglich durch. Was diese Cluster sind, und ob sie eine Entsprechung in der Realität haben, kann oftmals nur durch Heuristiken bestimmt werden.

2.3 Quellen der Opazität von KI-Systemen

Wie in Kapitel 1 dargestellt, ist die vorliegende Arbeit durch das Black-Box-Problem motiviert. Ein entscheidender Teil dieses Problems ist die epistemische Opazität vieler KI-Systeme, also die Schwierigkeit oder schlimmstenfalls Unmöglichkeit, diese Systeme gänzlich zu verstehen. Der vorangegangene Abriss über KI und ML hat die erforderlichen Grundlagen gelegt, um näher zu fassen, wie der spezifische Aufbau und die Funktionsweise eines KI-Systems zu dessen epistemischer Opazität führen können. Doch existieren darüber

hinaus noch weitere Opazitätsquellen, die zum Black-Box-Problem beitragen. In einem Aufsatz (Mann et al., 2023), den ich gemeinsam mit Barnaby Crook, Lena Kästner, Astrid Schomäcker und Timo Speith verfasst habe, haben wir bestehende Überlegungen zur Opazität von KI-Systemen und anderen Softwaresystemen zusammengeführt, erweitert und in Form einer Taxonomie von acht Quellen von Opazität systematisiert (siehe Abbildung 2.4). Dieser Abschnitt gibt Auszüge dieses Aufsatzes in überarbeiteter Form wieder.

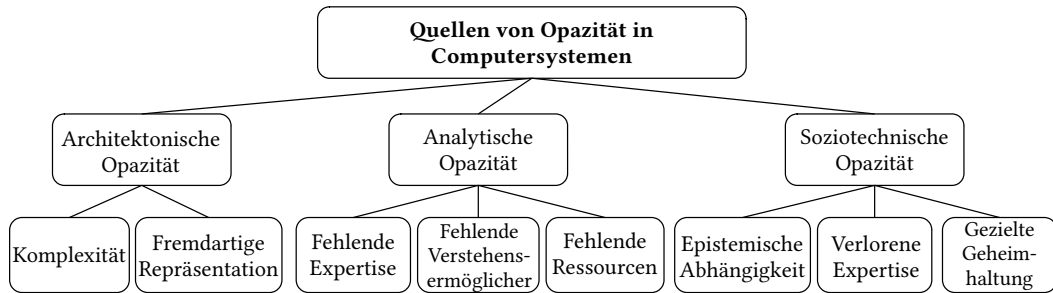


Abbildung 2.4: Die von mir und meinen Co-Autor:innen vorgeschlagene Taxonomie von Opazitätsquellen im Kontext von Computersystemen. Bildquelle: Eigene Darstellung, adaptiert von Mann et al. (2023). Siehe Teil IV für Details.

Meine Co-Autor:innen und ich haben die von uns identifizierten Opazitätsquellen in drei Gruppen eingeteilt: architektonische, analytische sowie soziotechnische Opazität. Bei der ersten Gruppe rührt die Opazität von der *Architektur* des KI-Systems. Diese Form der Opazität ist der Schwerpunkt der informatischen XAI-Forschung und ein wichtiger Grund für das Black-Box-Problem. Sie hängt vornehmlich mit den in den Abschnitten 2.1 und 2.2 geschilderten Eigenschaften verschiedener KI-Systeme zusammen. Beispielsweise gelten neuronale Netze aufgrund ihrer subsymbolischen Repräsentationen als besonders opak. In anderen Fällen besteht die Quelle von Opazität in unzureichenden materiellen oder epistemischen Voraussetzungen der Person, die das fragliche KI-System analysieren möchte. So macht etwa das individuelle Vorwissen einer Person einen Unterschied für die Verständlichkeit eines gegebenen KI-Systems. Solche Quellen sind der *analytischen* Opazität zugeordnet. Die dritte Gruppe umfasst *soziotechnische* Opazitätsquellen. Denn auch die sozialen oder institutionellen Kontexte, in die KI-Systeme eingebettet sind, können Opazität herbeiführen. Zum Beispiel kann ein proprietärer Algorithmus in erster Linie deshalb opak sein, weil er dem Geschäftsgeheimnis unterliegt.

Ich gehe nachfolgend näher auf jene Opazitätsquellen ein, die für die vorliegende Arbeit besonders relevant sind. In Abschnitt 2.3.1 betrachte ich die beiden Quellen architektonischer Opazität, die meine Co-Autor:innen und ich identifiziert haben: Komplexität und fremdartige Repräsentationen. In Abschnitt 2.3.2 gehe ich auf eine Auswahl von analytischen Opazitätsquellen ein: fehlende Expertise sowie fehlende Verstehensermöglicher. Schließlich erläutere ich in Abschnitt 2.3.3, wie epistemische Abhängigkeit – als eine der soziotechnischen Quellen von Opazität – Verstehen von KI-Systemen verhindern oder

erschweren kann.

2.3.1 Architektonische Opazität

Architektonische Opazität ergibt sich aus den strukturellen Eigenschaften von Computersystemen (ebd., S. 338). Zum einen können Computersysteme aufgrund ihrer Komplexität opak sein, zum anderen aufgrund der Repräsentationen, die sie verwenden.¹⁴

Komplexität

Die Komplexität von KI-Systemen wird häufig als Quelle von Opazität angeführt, doch bleibt meist offen, worin diese Komplexität besteht (vgl. auch Guidotti et al., 2019, S. 93:9). Je nach Disziplin können damit nämlich unterschiedliche Eigenschaften gemeint sein. Die informatische Komplexitätsforschung befasst sich mit der Einteilung algorithmisch behandelbarer Probleme in verschiedene Komplexitätsklassen. Komplexität wird dabei anhand des erforderlichen Rechenaufwands gemessen – sei es in Bezug auf KI-Modelle (siehe z. B. S. Russell & Norvig, 2021, S. 27, 45) oder Erklärbarkeitsansätze (Molnar, 2022, S. 23). Offensichtlich nützt ein KI-System oder ein Erklärbarkeitsansatz nichts, wenn der für die Nutzung erforderliche Rechenaufwand impraktikabel ist. Eine zu große computationale Komplexität lässt sich daher als Beispiel für analytische Opazität aufgrund fehlender Ressourcen auffassen (Mann et al., 2023, S. 338).

Das interdisziplinäre Forschungsfeld der Systemtheorie betrachtet dagegen verschiedene komplexe Systeme und versucht unter anderem, Komplexität in einem allgemeineren Sinne zu definieren. Eine vereinfachende, hier jedoch hilfreiche Unterscheidung innerhalb dieser Debatte identifiziert zwei grundsätzliche Komplexitätsdimensionen: die Anzahl der Elemente eines Systems und deren Interaktionen (Arbesman, 2017, S. 13–15).¹⁵

Die erste Komplexitätsdimension, also die Anzahl der Elemente, kann grundsätzlich alle Arten von KI-Systemen betreffen. Beispielsweise kann ein tiefer Entscheidungsbaum (oder eine seiner Entscheidungen) aufgrund der Anzahl der Knoten opak sein (Z. C. Lipton, 2018; van Otterlo, 2013). Bei einem neuronalen Netz mit Millionen oder gar Milliarden von Parametern¹⁶ verschärft sich diese Problematik natürlich weiter. Die zweite Dimension

¹⁴Architektonische Opazität kann auch dazu führen, dass das fragliche System nicht für die Erklärung oder Modellierung natürlicher Phänomene herangezogen werden kann, weil die Akkuratheit des KI-Systems, verstanden als Repräsentation eines Aspekts der Welt, nicht beurteilbar ist (Mann et al., 2023, S. 339; siehe dazu Facchini & Termine, 2022; Sullivan, 2022). Diese Variante der Opazität ist z. B. relevant, wenn Naturwissenschaftler:innen KI-Systeme im Rahmen ihrer Forschungstätigkeit nutzen.

¹⁵Was als distinktes Element eines Computersystems zählt, hängt vom Kontext ab. Informatiker:innen betrachten KI-Systeme vielleicht auf Ebene des Codes, der Neuronen oder gar der Hardware. Für Nichtfachleute kommen vornehmlich intuitiv verständliche Sinneinheiten als relevante Elemente in Frage (Mann et al., 2023, Fn. 2 auf S. 338). Zu weiteren Eigenschaften komplexer Systeme, wie etwa Emergenz oder dezentrale Organisation, siehe Davidsson et al., 2017; Ladyman et al., 2013.

¹⁶Bei künstlichen neuronalen Netzen sind Parameter jene numerischen Werte, die während des Trainings erlernt werden, also unter anderem die Kantengewichte (siehe Abschnitt 2.1.2).

besteht in der Interaktion der Elemente eines Systems. Bei vielen komplexen Systemen ist diese Interaktion nichtlinear (Davidsson et al., 2017, S. 784), was das Verhalten eines solchen Systems meist weniger intuitiv und schlechter vorhersagbar macht. Neuronale Netzwerke enthalten mit der Aktivierungsfunktion ein nichtlineares Element (siehe Abschnitt 2.1.2), doch auch konventionelle KI kann aufgrund der Interaktion ihrer Elemente komplex sein. Ein Beispiel sind so genannte ‚Kludges‘, also unelegante Lösungen für ein Programmierproblem, die (bis zu einem gewissen Punkt) zwar funktionieren, aber schlecht verstanden werden. Kludges entstehen oftmals aufgrund der sukzessiven, kollektiven (Weiter-)Entwicklung von Computerprogrammen und hängen daher eng mit soziotechnischen Opazitätsquellen zusammen (Mann et al., 2023, S. 338).

Ist ein System zu komplex, übersteigt es die kognitiven Kapazitäten eines Menschen. Man kann das System oder einen einzelnen seiner Entscheidungsprozesse nicht mental ‚simulieren‘ und in der Gänze erfassen, sondern bestenfalls jeden Schritt einzeln nachvollziehen (vgl. Z. C. Lipton, 2018, S. 12–13; Jebeile, 2018a, S. 73). Vor diesem Hintergrund ließe sich Komplexität auch als analytische Quelle von Opazität auffassen, insofern die menschliche Kognition der begrenzende Faktor ist (Mann et al., 2023, S. 338). Gleichmaßen können geeignete Verstehensmöglichkeiten die Funktionsweise oder den Entscheidungsprozess eines KI-Systems bestenfalls so herunterbrechen, dass auch komplexe Systeme ein Stück weit verständlich werden. Daher hängt diese Opazitätsquelle auch mit dem Fehlen geeigneter Verstehensmöglichkeiten zusammen.

Fremdartige Repräsentationen

Auch die Repräsentationen, die ein KI-System verwendet, können zu Opazität führen: entweder, weil die Art der Repräsentation für Menschen fremdartig ist, oder ihr Inhalt (ebd.). Beides betrifft in erster Linie Systeme, die auf distribuierten, subsymbolischen Repräsentationen beruhen (siehe Abschnitt 2.1.2).

Im ersten Fall korrespondieren die subsymbolischen Repräsentationen menschenverständlichen Begriffen, auch wenn diese Repräsentationen über eine große Anzahl von Neuronen verteilt und damit nicht unmittelbar epistemisch zugänglich sind. Beispielsweise deuten verschiedene Forschungsergebnisse darauf hin, dass manche neuronalen Netzwerke Repräsentationen erlernen, die Begriffen wie *Schnee*, *Haus* oder *Ofen* ähneln (Bau et al., 2020; Fong & Vedaldi, 2018). Werden diese nachvollziehbaren Sinneinheiten identifiziert, können sie für die Formulierung von Erklärungen verwendet werden. Allerdings lassen sich nicht alle Einheiten neuronaler Netze menschlichen Begriffen zuordnen, dieselbe Einheit kann mehrere Begriffe erkennen, und derselbe Begriff kann von mehreren (Kombinationen von) Einheiten erlernt werden (Freiesleben, 2026, Kap. 13.5; Molnar, 2022, Kap. 9.1.2). Zudem scheinen die fraglichen Einheiten häufig nicht dieselbe funktionale Rolle wie die zugehörigen menschlichen Begriffe einzunehmen – das Entfernen der Einheiten reduziert die Systemperformance oftmals kaum (Freiesleben, 2026).

Im zweiten Fall ist auch der Inhalt der Repräsentation für Menschen fremdartig (vgl.

Facchini & Termine, 2022, S. 85–86). Beispielsweise legt die Forschung zu ‚feindlichen Beispielen‘ (engl. *adversarial examples*; Szegedy et al., 2014) nahe, dass manche KI-Systeme Informationen radikal anders ‚wahrnehmen‘ und verarbeiten als Menschen. Feindliche Beispiele werden insbesondere im Kontext von modernen Bildklassifikatoren diskutiert. Es handelt sich dabei um gezielt generierte oder ausgewählte Bilder, die vom System entgegen menschlicher Intuitionen falsch klassifiziert werden. Klassischerweise erhält man feindliche Beispiele durch die Veränderung eines natürlichen Bildes, das mit hoher Wahrscheinlichkeit richtig klassifiziert wird.¹⁷ Nach der subtilen Veränderung erscheint das Bild für menschliche Betrachter:innen oft unverändert, erhält aber vom KI-System eine andere, scheinbar willkürliche Klassifizierung. Die Existenz feindlicher Beispiele spricht dafür, dass einige Systeme auf subtile Muster reagieren, die für Menschen keinen Sinn ergeben (Mann et al., 2023, S. 338–339; vgl. auch Mueller, 2020; Pelland et al., 2024).

Im Kontext der Verwendung neuronaler Netze in der Wissenschaft schlägt Florian Boge angesichts dessen vor, dass diese Systeme nicht etwa Begriffe, sondern ‚funktionale Begriffs-Proxies‘ erlernen. Diese Proxies nehmen in einigen, aber nicht notwendigerweise allen Kontexten dieselbe Rolle ein wie Begriffe. Allerdings können diese Proxies prinzipiell gut generalisierbar und daher fruchtbar für wissenschaftliche Entdeckungen und Vorhersagen sein, auch wenn sie keinen menschlichen Begriffen entsprechen (Boge, 2024). Auch Andrew Ilyas et al. argumentieren, dass solche für Menschen unverständlichen Muster durchaus informativ seien und es sich daher bei feindlichen Beispielen keineswegs um einen Hinweis auf Fehler handle, die behoben werden müssen (Ilyas et al., 2019). Dieser Interpretation zufolge wenden bestimmte KI-Systeme mitunter Strategien an, die zwar erfolgreich sind, aber radikal von den (bewussten) Strategien von Menschen abweichen und aufgrund dieser Andersartigkeit opak sind.

2.3.2 Analytische Opazität

Analytische Opazität ergibt sich durch das Fehlen der Expertise oder der Verstehensermöglicher, die erforderlich wären, um das System zu analysieren und damit transparent zu machen. Wie in Abschnitt 2.3.1 erwähnt, können darüber hinaus auch unzureichende Ressourcen zu analytischer Opazität führen (mehr dazu in Mann et al., 2023, S. 339).

¹⁷Es gibt diverse Möglichkeiten, feindliche Beispiele zu erhalten (einen Überblick bietet Molnar, 2022, Kap. 9.4). Beispielsweise können ‚feindliche Aufkleber‘ (Brown et al., 2017), die auf realen Objekten (etwa Verkehrsschildern) angebracht werden, jede beliebige Klassifikation dieses Objekts herbeiführen (z. B. durch autonome Fahrzeuge). Es gibt auch ‚natürliche feindliche Beispiele‘. Dabei handelt es sich um Bilder, die ohne gezielte Manipulation von verschiedenen Klassifikationssystemen mit einer sehr hohen Wahrscheinlichkeit falsch klassifiziert werden – einfach, weil sie sich in relevanten Hinsichten von den Trainingsdaten unterscheiden (Hendrycks et al., 2021).

Fehlende Expertise

Wie viele Autor:innen betonen, ist es plausibel, dass Opazität mitunter von individuellen Vorkenntnissen abhängt (Burrell, 2016; Facchini & Termine, 2022; Jebeile, 2018a). Beispielsweise können fehlende Programmierkenntnisse, mangelndes Fachwissen über KI-Systeme oder die Anwendungsdomäne dazu führen, dass ein gegebenes KI-System für einen Laien opak ist, für eine Expertin hingegen nicht (Mann et al., 2023, S. 337).

Diese Quelle von Opazität gewinnt an Bedeutung, da KI-Systeme zunehmende Verbreitung in diversen Anwendungskontexten finden und mittlerweile unterschiedlichste Gruppen von Stakeholdern, darunter auch viele Nichtfachleute, mit KI arbeiten. Allerdings können auch KI-Expert:innen von Opazität aufgrund fehlender Expertise betroffen sein. Beispielsweise sind die meisten Programmierer:innen lediglich mit so genannten höheren Programmiersprachen vertraut, nicht aber mit den zugrundeliegenden Maschinensprachen (Creel, 2020, S. 578). Entsprechend fehlt auch einer Informatikerin üblicherweise das Fachwissen, um ein gegebenes KI-System lückenlos zu verstehen.

Die kognitiven und zeitlichen Kapazitäten eines Individuums, und damit seine Möglichkeiten, sich Kenntnisse oder Fähigkeiten anzueignen, sind begrenzt. Daher ist diese Quelle oftmals mit epistemischer Abhängigkeit von Personen mit abweichender Expertise (siehe unten) verknüpft (Mann et al., 2023, S. 339). Außerdem kann fehlende Expertise auch als das Fehlen eines geeigneten Verstehensermöglicher aufgefasst werden, der die Wissens- bzw. Verstehenslücken des fraglichen Individuums kompensiert. Beispielsweise werden XAI-Methoden allzu oft ‚von Informatiker:innen für Informatiker:innen‘ entwickelt und berücksichtigen andere relevante Zielgruppen von Erklärbarkeitsansätzen nur unzureichend (Miller et al., 2017).

Fehlende Verstehensermöglicher

Wie die umfangreiche Debatte um Erklärbarkeitsmethoden zeigt (siehe dazu Abschnitt 3.2), kann Transparenz mitunter lediglich durch die Verwendung bestimmter Verstehensermöglicher erzielt werden. Sind diese situativ nicht verfügbar oder womöglich nicht einmal entwickelt worden, bleibt ein System für einen Stakeholder opak (Mann et al., 2023, S. 339).¹⁸ Neben XAI-Ansätzen lassen sich auch Tools wie Datenbanken, themenspezifische Internetforen (z. B. *Stack Overflow*), höhere Programmiersprachen (wie *Python* oder *Java*) oder Entwicklungsplattformen (etwa *GitHub*) als Verstehensermöglicher auffassen, deren Fehlen ein Computersystem für ein Individuum opak machen kann (ebd.).

Häufig kann diese Quelle von Opazität auch als das Fehlen von Werkzeugen zur Überwindung anderer Opazitätsquellen verstanden werden (ebd.). Zum Beispiel sind XAI-Methoden Verstehensermöglicher, welche die architektonische Opazität von KI-Systemen adressieren – etwa, indem ein Erklärbarkeitsansatz die Komplexität des systeminternen Entscheidungs-

¹⁸Ich gehe in Kapitel 3 näher darauf ein, was Verstehensermöglicher sind, und Kapitel 12 arbeitet einige Eigenschaften besonders vorteilhafter Verstehensermöglicher heraus.

prozesses auf eine kleine Auswahl von Faktoren herunterbricht. Und Datenanalyst:innen verwenden verschiedene Visualisierungen (etwa Histogramme oder Graphen), um sich einen Überblick über die teils großen Mengen an Trainingsdaten von ML-basierten Systemen zu verschaffen (vgl. Skiena, 2017, Kap. 6) oder die von wissenschaftlichen Computersimulationen generierten Daten zu analysieren (Jebeile, 2018b) – also der Komplexität der Daten Herr zu werden.

2.3.3 Soziotechnische Opazität

Soziotechnische Opazität ergibt sich aus den sozialen, gesellschaftlichen und technologischen Strukturen, in die ein Computersystem eingebunden ist (Mann et al., 2023, S. 339). Im Kontext dieser Arbeit ist vornehmlich Opazität durch epistemische Abhängigkeit interessant, daher beschränke ich mich hier auf diese Quelle. In unserem Aufsatz gehen meine Co-Autor:innen und ich darüber hinaus auf Opazität durch verlorenes Wissen und gezielte Geheimhaltung ein (ebd., S. 340).

Epistemische Abhängigkeit

Die Entwicklung moderner KI- bzw. Computersysteme beruht auf der Arbeitsteilung zwischen unterschiedlichen Spezialist:innen, die jeweils verschiedene Komponenten des Systems beisteuern. Denn eine einzelne Entwicklerin besitzt nur einen Bruchteil der Expertise, die erforderlich wäre, um ein solches System von Grund auf zu kreieren – dazu zählt zum Beispiel auch der Bau der Hardware oder die Entwicklung der Programmiersprache, in der das fragliche Computerprogramm geschrieben wird. Und selbst eine Programmiererin nutzt häufig wiederverwendbare, von anderen bereitgestellte Komponenten (z. B. Programmbibliotheken), ohne selbst ein tieferes Verständnis davon zu besitzen (ebd.; siehe dazu Mittelstadt et al., 2016, S. 10 und Endnote 8 auf S. 15). Kurz: KI-Systeme sind das Ergebnis der Arbeitsteilung zwischen Personen mit unterschiedlicher Expertise, die jeweils (epistemisch) abhängig voneinander sind und für die der Beitrag ihrer Kolleg:innen wenigstens teilweise opak ist.

Susann Wagenkecht diskutiert das Phänomen der epistemischen Abhängigkeit im Kontext der wissenschaftlichen Produktion von Wissen (Wagenknecht, 2014). Ihr Vorschlag lässt sich aber auch auf die kollaborative Entwicklung von Computer- bzw. KI-Systemen übertragen. Für Wagenknecht ist eine Wissenschaftlerin *A* in *opaker Weise* epistemisch abhängig von der Arbeit von Kollegin *B*, wenn *A* ihre Überzeugung, dass *p*, einzig durch Bezug auf *B* rechtfertigen kann – und zwar, weil *A* nicht die nötige Expertise besitzt, um die Wahrheit dieser Überzeugung durch eigene Beobachtungen und/oder Schlussfolgerungen festzustellen. Folglich muss sich *A* auf das Zeugnis von *B* verlassen (ebd., S. 484). Anders ausgedrückt, besitzt *A* nicht die erforderliche Expertise, um den Beitrag von *B* zur gemeinsamen Aufgabe unabhängig zu bewerten oder selbst zu leisten (ebd., S. 483). *Transluzente* epistemische Abhängigkeit besteht hingegen dann, wenn *A* die erforderliche

Expertise besitzt, um unabhängig von *B* zu arbeiten, sich aber aus pragmatischen Gründen dagegen entscheidet (Wagenknecht, 2014, S. 485). Beide Formen der epistemischen Abhängigkeit führen zu einem gewissen Maß an Opazität (daher spricht Wagenknecht auch von *transluzenter* statt *transparenter* epistemischer Abhängigkeit, ebd., S. 485–486). Allerdings ist opake epistemische Abhängigkeit schwieriger zu überwinden, da dies den aufwändigen Erwerb von zusätzlicher Expertise erfordert (siehe Abschnitt 2.3.2).

Epistemische Abhängigkeit kann zum Problem werden. Die interdependente, auf der Kollaboration verschiedenster Akteur:innen beruhende Entwicklung von (auch systemkritischen) Computersystemen erschwert beispielsweise das Auffinden von Sicherheitslücken, da es niemanden gibt, der alle Systemkomponenten vollständig überblickt. Dies lässt sich an der im Jahr 2018 bekannt gewordenen Anfälligkeit von modernen Computerprozessoren, also einem Teil der Hardware, gegenüber so genannten *Spectre*- bzw. *Meltdown*-Angriffen erkennen. Bis dahin war Cybersicherheit in erster Linie die Domäne von Software-Entwickler:innen, während sich bei der Entwicklung von Hardware vornehmlich auf die Steigerung der Leistungsfähigkeit konzentriert wurde (Abu-Ghazaleh et al., 2019, S. 49). Eine starke epistemische Arbeitsteilung zwischen Hardware-Entwicklung und (software-fokussierter) Cybersicherheit birgt angesichts solcher Schwachstellen ein potenzielles Sicherheitsrisiko.

Doch Opazität aufgrund von epistemischen Abhängigkeiten hat auch positive Seiten: Selbst wenn ein gegebenes KI-System für eine Einzelperson bis zu einem gewissen Grad stets opak bleiben mag, kann die Kollaboration verschiedener, epistemisch abhängiger Expert:innen insgesamt betrachtet die Opazität von Computersystemen reduzieren (Mann et al., 2023, S. 340). In diesem Sinne ist epistemische Abhängigkeit nicht nur eine Quelle von Opazität, sondern auch von Transparenz.

Wir halten fest: KI-Systeme können aus verschiedenen Gründen epistemisch opak sein. Wie in Kapitel 1 geschildert, kann diese Opazität der Erfüllung gesellschaftlicher Desiderata im Wege stehen. Im nun folgenden Kapitel geht es daher um Strategien zur Reduzierung der Opazität von KI-Systemen – insbesondere solcher Opazität, die durch architektonische Quellen bedingt ist.

3 Erklärbarkeit

Die Erklärbarkeit von KI-Systemen ist kein neues Thema, sondern wird spätestens seit einem der ersten erfolgreichen Expertensysteme, MYCIN (Shortliffe et al., 1973), und dem darauffolgenden Boom solcher Systeme diskutiert (Clancey, 1983; Swartout & Moore, 1993). Während Expertensysteme zu den symbolischen KI-Systemen gehören und daher wenigstens für Expert:innen recht gut verständlich sind, erschweren es die subsymbolischen Repräsentationen neuronaler Netze, ihre Funktionsweise nachzuvollziehen (siehe Abschnitt 2.3.1). Im Zuge des gegenwärtigen Erfolgs letzterer Systeme hat die informatische Forschung zu Erklärbarkeit (mittlerweile häufig unter dem Namen ‚Explainable AI‘ oder kurz ‚XAI‘) in den letzten Jahren massiv zugenommen (Nauta et al., 2023, Abb. 4 auf S. 13).

Erklärbarkeit ist ein wichtiger Weg, um KI-Systeme oder einzelne ihrer Aspekte (instrumentell) verständlich zu machen. Dieses Kapitel stellt daher die nötigen informatischen und philosophischen Grundlagen zu Erklärbarkeit zusammen.¹⁹ In Abschnitt 3.1 erläutere ich das hier zugrunde gelegte Verständnis von Erklärbarkeit und verwandter Begriffe. In Abschnitt 3.2 systematisiere ich Erklärbarkeitsmethoden anhand einiger zentraler Kategorien der Debatte und gehe auf die Evaluation von XAI-Erklärungen ein. In Abschnitt 3.3 gehe ich der Frage nach, ob KI-Systeme im philosophischen Sinne des Wortes überhaupt erklärbar sind und ob Erklärungen der einzige Weg sind, um Aspekte von KI-Systemen verständlich zu machen.

3.1 Was ist Erklärbarkeit?

Die Klärung des Begriffs *Erklärbarkeit* wird dadurch erschwert, dass innerhalb der XAI-Forschung Uneinigkeit über seine exakte Bedeutung und jene verwandter Begriffe besteht (Krishnan, 2020; Z. C. Lipton, 2018; Longo et al., 2024, Kap. 3.3). Insbesondere informatische Aufsätze führen eine Vielzahl semantisch ähnlicher (englischer) Termini an, etwa ‚comprehensibility‘, ‚explainability‘, ‚explicability‘, ‚intelligibility‘, ‚interpretability‘, ‚perspicuity‘, ‚transparency‘ oder ‚understandability‘. In den meisten Aufsätzen werden diese und ähnliche Ausdrücke ohne weitere Erläuterung verwendet. In jüngerer Zeit wurde sich zwar auch in der Informatik vermehrt um Definitionen bemüht,²⁰ allerdings tragen

¹⁹Die Arbeit an diesem Kapitel hat maßgeblich vom wertvollen Feedback von Lena Kästner profitiert. Außerdem möchte ich mich bei Timo Speith für die kritische Prüfung der informatischen Aspekte bedanken. Wie im vorherigen Kapitel gilt natürlich auch hier, dass etwaige Fehler meine eigenen sind.

²⁰Eine Übersicht über verschiedene Charakterisierungen von einigen zentralen Begriffen und deren Verhältnis bietet Timo Speiths Dissertation (Speith, 2023, S. 127–130).

diese Definitionen meinem Eindruck nach nicht immer zur Klärung der jeweiligen Begriffe bei. Dies lässt sich exemplarisch anhand der Übersicht über typische XAI-Terminologie von Sina Mohseni et al. illustrieren. Die Autor:innen versuchen, gängige Bedeutungen relevanter Ausdrücke zu synthetisieren und zu systematisieren. Zu diesem Zweck setzen sie *intelligible system* und *transparent AI* als übergeordnete Begriffe fest und ordnen diesen beiden Begriffen verschiedene Unterbegriffe zu. Hier sind einige der Beschreibungen, die Mohseni et al. zur Erläuterung der von ihnen systematisierten Begriffe angeben:

- *Understandability/Intelligibility*: „Intelligible systems support user understanding of [the] system’s underlying functions“ (Mohseni et al., 2020, Tab. 1 auf S. 1:8)
- *Interpretability*: „The ability to support user understanding and comprehension of the model decision making process and predictions“ (ebd.)
- *Explainability*: „The ability to explain [the] underlying model and its reasoning with accurate and user comprehensible explanations“ (ebd.)
- *Explainable AI*: „Supporting user understanding of complex models by providing explanations for predictions“ (ebd.)

„Understandability/intelligibility“ ordnen Mohseni et al. (neben anderen Ausdrücken) der Kategorie *intelligible system* zu. „Explainable AI“, „interpretability“ und „explainability“ werden hingegen *transparent AI* zugewiesen.²¹ Meinem Eindruck nach sind die obigen Beschreibungen charakteristisch für die Begriffsarbeit in der informatischen XAI-Literatur. Nicht nur wird die Unterscheidung zwischen „transparent system“ und „intelligible system“ kaum erläutert. Es bleibt auch offen, wie sich die Termini „understanding“/„comprehension“ oder „decision making process“/„reasoning“/„underlying functions“ zueinander verhalten und weshalb „understandability/intelligibility“, im Gegensatz zu „interpretability“ oder „explainability“, keine Fähigkeit ist. Mohseni et al. berufen sich auf bestehende Redeweisen in der XAI-Literatur und bleiben eine Antwort auf diese Fragen schuldig.

Das Hauptproblem mit diesen und ähnlichen Definitionen besteht aber darin, dass ihr Mehrwert nicht deutlich wird. Ich bezweifle nicht, dass sich einige hilfreiche Unterscheidungen zwischen solchen Ausdrücken treffen ließen. Weniger überzeugt bin ich davon, dass es sinnvoll und möglich ist, im XAI-Kontext trennscharf zwischen einem guten dutzend Termini zu differenzieren, deren Bedeutungen sich offensichtlich stark überschneiden. Ich möchte in dieser Arbeit daher den XAI-bezogenen Begriffsapparat möglichst übersichtlich halten. Weil sich mein Vorhaben im Kern um Verstehen dreht, ist es attraktiv, Erklärbarkeit

²¹Zusätzlich unterscheiden Mohseni et al. zwischen praktischen (informatischen) Ansätzen, erwünschten Eigenschaften von KI-Systemen und erwünschten Ergebnissen (die verschiedenen gesellschaftlichen Desiderata entsprechen). „Understandability/intelligibility“, „explainability“ und „interpretability“ sind erwünschte Eigenschaften, „explainable AI“ bezeichnet einen praktischen Ansatz (Mohseni et al., 2020, Tab. 1 auf S. 1:8).

durch Bezugnahme auf diese epistemische Errungenschaft ausbuchstabieren. Auch viele Charakterisierungen von Erklärbarkeit, Interpretierbarkeit oder XAI sprechen für diese Herangehensweise:

Different notions of rendering opaque AI approaches understandable in some sense go by the terms “explainable,” “interpretable,” “intelligible,” and “transparent,” to name the dominant ones. (Herzog, 2022, S. 50)

Since there are a lot of different approaches for solving this [black-box] problem, each with their own individual needs and goals, there is no single common definition of the term XAI. The key idea, however, is to enable users to understand the decision-making of any model. (Burkart & Huber, 2021, S. 247)

[E]xplainability or interpretability of a model measures the degree to which a human observer can understand the reasons behind a decision (e.g. a prediction) made by the model. (Dam et al., 2018, S. 54)

In general, the goal of explainability is to make certain aspects of a system understandable for humans, being these developers, designers, regulators or users from general society. (Longo et al., 2024, S. 3; Literaturangabe entfernt)

Interpretable Machine Learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans. (Molnar, 2022, S. 11)

The XAI program’s goal is to create a suite of new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems. (Gunning & Aha, 2019, S. 45)

Die obigen Zitate illustrieren nicht nur die unterschiedlichen Auffassungen von Begriffen wie *Erklärbarkeit*, *Interpretierbarkeit* oder *XAI*. Sie legen auch nahe, dass die Bezugnahme auf Verstehen einen möglichen Minimalkonsens innerhalb der XAI-Literatur darstellt (vgl. Arrieta et al., 2020, S. 85; Fleisher, 2022). Dieser Minimalkonsens bildet den Ausgangspunkt dafür, wie ich relevante Ausdrücke im Rahmen dieser Arbeit verwenden möchte:

- *(epistemische) Opazität/Transparenz*: Ein transparentes KI-System bzw. ein transparenter Aspekt dieses Systems ist verständlich insofern, dass keinerlei Opazitätsquellen dem Erwerb von Verstehen im Wege stehen. Analog dazu ist ein opakes System (oder ein opaker Aspekt) unverständlich, wenn eine oder mehrere Opazitätsquellen zum Tragen kommen (siehe Abschnitt 2.3). Zwischen diesen beiden Extremen liegen Grade der Opazität, die mit einem gewissen Verstehensgrad vereinbar sind.
- *Erklärbarkeit* ist die Eigenschaft eines KI-Systems (oder einzelner seiner Aspekte), dank des Vorhandenseins einer geeigneten Erklärung oder eines anderen Verstehensermöglichers verständlich (also transparent) zu sein. Verwandte Begriffe – wie *Interpretierbarkeit* bzw. *interpretability*, *intelligibility*, *perspicuity* oder *explicability* – subsumiere ich nachfolgend ebenfalls unter ‚Erklärbarkeit‘.

- *XAI- bzw. Erklärbarkeitsforschung* verwende ich als Sammelbegriff für das interdisziplinäre Forschungsfeld, das sich in theoretischer oder praktischer Hinsicht damit beschäftigt, wie (Aspekte von) KI-Systemen, die von aus unterschiedlichen Quellen herrührender Opazität betroffen sind, für Menschen verständlich gemacht werden können oder sollten. Im engeren (innerhalb der Informatik üblichen) Sinne umfasst XAI ausschließlich informatische bzw. ingenieurwissenschaftliche Forschung, die technische Lösungen zur Überwindung oder Vermeidung von vornehmlich architektonischer Opazität (siehe Abschnitt 2.3.1) anstrebt.
- *XAI-Methoden bzw. Erklärbarkeitsansätze* sind technische Strategien, die darauf ausgerichtet sind, vornehmlich die architektonische Opazität von (Aspekten von) KI-Systemen zu überwinden und diese dadurch für unterschiedliche Stakeholder verständlich zu machen. Solche Erklärbarkeitsmethoden gehören XAI im engen Sinne an und werden in Abschnitt 3.2 näher vorgestellt.
- *Verstehensermöglicher*: Verstehensermöglicher sind Repräsentationen systembezogener Aspekte, die explanatorisch sein können, aber nicht müssen. Es handelt sich um einen Oberbegriff, der neben (XAI-)Erklärungen auch weitere Mittel des Verstehenserwerbs einschließt (z. B. das Zeugnis einer Expertin, technische Dokumentationen oder nicht-propositionale²² Repräsentationen). Zudem adressieren Verstehensermöglicher Opazität, die von unterschiedlichsten, nicht nur architektonischen, Quellen herrührt. Ich charakterisiere Verstehensermöglicher in Abschnitt 3.3 genauer.

3.2 Informatische Erklärbarkeitsmethoden

Erklärbarkeit wird, wie oben erwähnt, häufig von einem informatischen Standpunkt aus betrachtet: Es sollen technische Lösungen für die architektonische, letztlich technologisch bedingte, Opazität von KI-Systemen gefunden werden. Zwar nehme ich in dieser Arbeit eine breitere Perspektive ein und fasse unter dem Begriff *Verstehensermöglicher* unterschiedliche Repräsentationen zusammen, die über die Ausgaben gängiger XAI-Methoden hinausgehen. Nichtsdestotrotz werden nachfolgend einige Grundlagen zu XAI im engeren, informatischen Sinne vorausgesetzt. Dieser Abschnitt bietet einen Überblick.

²²Für meine Zwecke genügt eine Standardauffassung von Propositionen als primäre Träger von Wahrheitswerten und Objekt von u. a. Überzeugungen. Konkreter können wir ein Fregeanisches Verständnis annehmen, demnach eine Proposition (ein ‚Gedanke‘) eine Kombination von abstrakten, begrifflich-deskriptiven Entitäten (oder ‚Sinnen‘) ist, die unabhängig vom Mentalen existiert und sprachlich unterschiedlich manifestiert werden kann. Entsprechend können mehrere Subjekte zugleich glauben, dass p , und p existiert unabhängig davon, ob jemand p erwägt. Fregeanische Propositionen sind feinkörnig, insofern unterschiedliche ‚Sinne‘ unterschiedliche Propositionen konstituieren: Sätze wie ‚Ich habe Locken‘ drücken je nach Sprecher:in verschiedene Propositionen mit potenziell abweichenden Wahrheitswerten aus. Auch die Substitution eines Ausdrucks durch einen co-referenziellen Ausdruck resultiert in einer neuen Proposition mit potenziell abweichendem Wahrheitswert – das von Gottlob Frege (1892) adaptierte Standardbeispiel betrifft die Substitution der co-referenziellen Ausdrücke ‚Venus‘ und ‚Abendstern‘.

War Erklärbarkeit vor einigen Jahren eher ein Nischenthema, ist die Anzahl der Methoden mittlerweile längst unüberschaubar geworden. Verschiedene Autor:innen haben sich bemüht, der Publikationsflut durch unterschiedliche Taxonomien und sonstige Formen der Systematisierung Herr zu werden (zum Beispiel Angelov et al., 2021; Arrieta et al., 2020; Belle & Papantonis, 2021; Guidotti et al., 2019; Nauta et al., 2023; Speith, 2022). Um grundlegende Herangehensweisen an Erklärbarkeit darzustellen, ist eine, wie Timo Speith es nennt, *begriffliche* Systematisierung anhand intuitiv zugänglicher Kategorien hilfreich (Speith, 2022, S. 2241; siehe Abbildung 3.1).

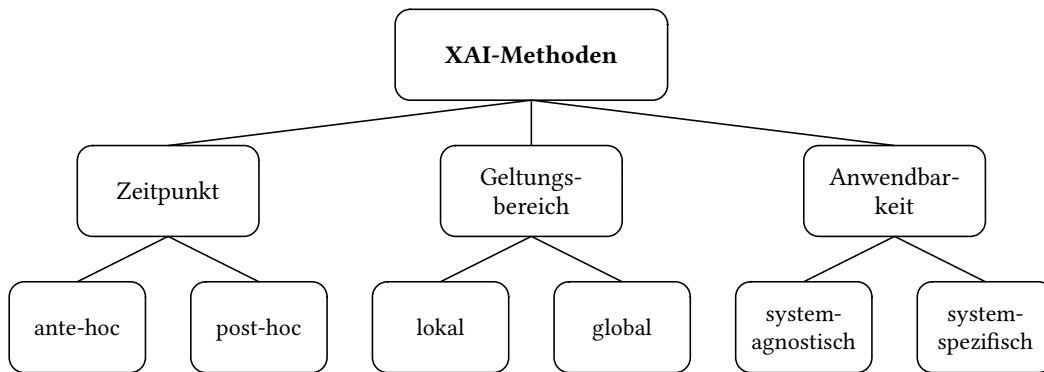


Abbildung 3.1: Die hier zugrunde gelegte begriffliche Taxonomie von XAI-Methoden. Nicht berücksichtigt werden mögliche Interdependenzen zwischen diesen Kategorien. Beispielsweise wird die Ausdifferenzierung der Anwendbarkeit und des Geltungsbereichs mitunter nur auf Post-hoc-Methoden bezogen. Bildquelle: Eigene Darstellung, adaptiert von Speith (2022). Siehe Teil IV für Details.

In Abschnitt 3.2.1 betrachte ich die erste dieser Kategorien, nämlich den Zeitpunkt, zu dem die Methode angewandt wird. Die zweite Kategorie betrifft den Geltungsbereich der generierten Erklärung (Abschnitt 3.2.2). Die dritte und letzte Kategorie systematisiert XAI-Ansätze gemäß ihrer Anwendbarkeit auf verschiedene KI-Systeme (Abschnitt 3.2.3). Abschließend gehe ich in Abschnitt 3.2.4 auf Evaluationskriterien ein, anhand derer die Güte einer XAI-Erklärung bestimmt werden kann.

3.2.1 Zeitpunkt: ante-hoc versus post-hoc

Die erste begriffliche Unterscheidung von XAI-Ansätzen betrifft den Zeitpunkt, zu dem Erklärbarkeit greift. Entscheidend dafür ist der Typ des KI-Systems. Vor allem Systeme, die auf symbolischen Repräsentationen beruhen, gelten als ante-hoc erklärbar (mitunter werden diese Systeme auch als ‚transparent‘ oder ‚inhärent erklärbar‘ bezeichnet). Sie sind, zumindest für Expert:innen, *qua* Design und damit von Anfang an relativ gut verständlich. Post-hoc erklärbare Systeme hingegen müssen durch aufwändigere Erklärbarkeitsansätze nachträglich verständlich gemacht werden.

Ante-hoc erklärbare Systeme

Es gibt eine Reihe von Systemtypen, die üblicherweise als ante-hoc erklärbar angesehen werden (siehe etwa Arrieta et al., 2020, Kap. 3; Belle & Papantonis, 2021, Kap. 6; Burkart & Huber, 2021, Kap. 3.3; Guidotti et al., 2019, Kap. 3.3). Dazu zählen unter anderem lineare Modelle, Entscheidungsbäume (siehe Abschnitt 2.1.1), regelbasierte Systeme (wie z. B. manche Expertensysteme, siehe Abschnitt 2.2.1) oder k -nächste-Nachbarn-Systeme (engl. *k-nearest neighbors*, k NN).

Die Einstufung eines KI-Systems als ante-hoc erklärbar stellt nicht sicher, dass das System auch tatsächlich transparent im oben beschriebenen Sinne ist. Das Label ist eher als Hinweis darauf zu verstehen, dass die vom System verwendeten Repräsentationen normalerweise für Menschen verständlich sind, dass also eine für subsymbolische Systeme charakteristische Quelle von Opazität vermieden wird (siehe Abschnitt 2.3.1). Dass dies nicht ausreicht, um ein KI-System verständlich zu machen, ist kein Geheimnis (Z. C. Lipton, 2018; van Otterlo, 2013, S. 57): Sehr tiefe oder breite Entscheidungsbäume – oder gar durch Ensemble-Methoden generierte *Random Forests*, die aus einer großen Anzahl von Entscheidungsbäumen bestehen und per Mehrheitsvotum eine Ausgabe generieren – sind meistens zu komplex, um ohne die nachträgliche Anwendung von Erklärbarkeitsmethoden verständlich zu sein (Arrieta et al., 2020, S. 91; Belle & Papantonis, 2021, S. 8). Gleiches gilt für Expertensysteme, die eine große Anzahl von komplexen Regeln verwenden (Arrieta et al., 2020, S. 91), oder k NN-basierte Systeme, die auf einer komplizierten Distanzfunktion basieren (Belle & Papantonis, 2021, S. 7).

Selbst weniger komplexe Vertreter ante-hoc erklärbarer Systeme sind für sich genommen oft nicht hinreichend transparent. Der Einfluss von Aspekten der Eingabedaten, die ein lineares Modell heranzieht, muss zumindest für Nichtfachleute erst visualisiert werden, um verständlich zu sein (Arrieta et al., 2020, S. 88). Und dass die Funktionsweise eines Expertensystems für die meisten Stakeholder nicht verständlich wird, indem man einfach die Schritte wiedergibt, die darin ablaufen, ist spätestens seit den 1980ern bekannt (Clancey, 1983; Swartout & Moore, 1993). Letztlich bedürfen also auch so genannte ante-hoc erklärbare Systeme zusätzlicher Verstehensmöglichkeiten, um für relevante Stakeholder verständlich zu sein. Üblicherweise werden aber nur solche Erklärbarkeitsansätze als Post-hoc-Methoden klassifiziert, die auf subsymbolische KI-Systeme angewandt werden.

Post-hoc-Methoden

Mittlerweile gibt es eine unüberschaubare Anzahl solcher Post-hoc-Ansätze. Exemplarisch sei eine wichtige Gruppe dieser Methoden genannt, die Attributionsmethoden. Diese Methoden ‚attribuieren‘ Features der Eingabedaten einen Wert, der den Einfluss auf die Ausgabe ausdrückt.²³ Intuitiv zugänglich ist dies bei der Klassifikation von Bildern: At-

²³Dazu zählen beispielsweise LIME (*Local Interpretable Model-Agnostic Explanations*; Ribeiro et al., 2016), LRP (*Layer-wise Relevance Propagation*, Bach et al., 2015), Grad-CAM (*Gradient-weighted Class Activa-*

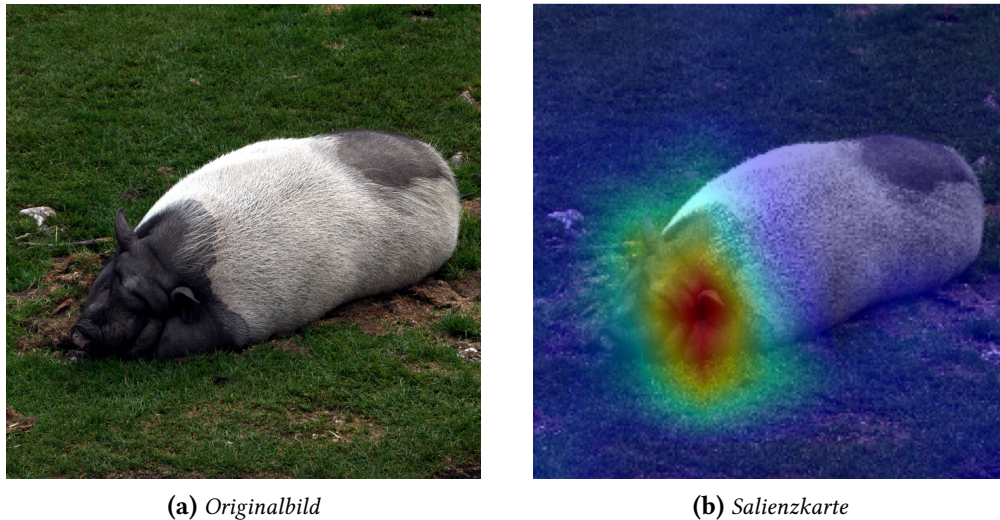


Abbildung 3.2: Grad-CAM-Erklärung: Zu sehen ist das Originalbild (3.2a) sowie die visuelle Erklärung (3.2b) für dessen Klassifikation als ‚Schwein‘ durch das Modell ResNet-50 (77,28%). Von rot hin zu blau eingefärbten Bildbereichen nimmt der Einfluss auf die Klassifikation ab. Offenbar war der Ohrbereich am wichtigsten für die Klassifikation. Bildquellen: Privat / eigene Darstellung unter Verwendung von Grad-CAM (Selvaraju et al., 2017). Siehe Teil IV für Details.

tributionsmethoden berechnen den Einfluss von einzelnen Pixeln oder Pixelgruppen (so genannten ‚Superpixeln‘) auf die finale Klassifikation eines Bildes. Den Einfluss verschiedener Bildbereiche auf die Ausgabe visualisieren Attributionsmethoden dann in der Regel durch Salienzkarten (engl. *saliency maps*).²⁴ Typischerweise wird dazu die Stärke des Einflusses auf die Klassifikation farbkodiert und dann halbtransparent als zusätzliche Bildebene in das ursprüngliche Bild eingefügt oder als ergänzende Visualisierung bereit gestellt (siehe Abbildung 3.2b), alternativ können weniger einflussreiche Bildbereiche auch gänzlich ausgeblendet werden (siehe Abbildung 3.3b).

tion Mapping; Selvaraju et al., 2017) oder SHAP (*Shapley Additive Explanation*; Lundberg & Lee, 2017). Die technischen Hintergründe sind komplex und unterscheiden sich teils stark. SHAP basiert auf spieltheoretischen Annahmen; LIME approximiert das zu erklärende KI-System lokal durch ein ante-hoc erklärbares lineares Modell; Grad-CAM berechnet den Gradienten der Klasse und LRP verfolgt das ‚Signal‘, das bei der Klassifikation durchs Netzwerk wandert, rückwärts von der Ausgabeschicht zurück zu den Eingabe-Features.

²⁴In der Literatur kursieren unterschiedliche verwandte Termini, beispielsweise ‚Heatmap‘ oder ‚Occlusion Map‘. Ich nutze der Einfachheit halber Salienzkarte als Oberbegriff.

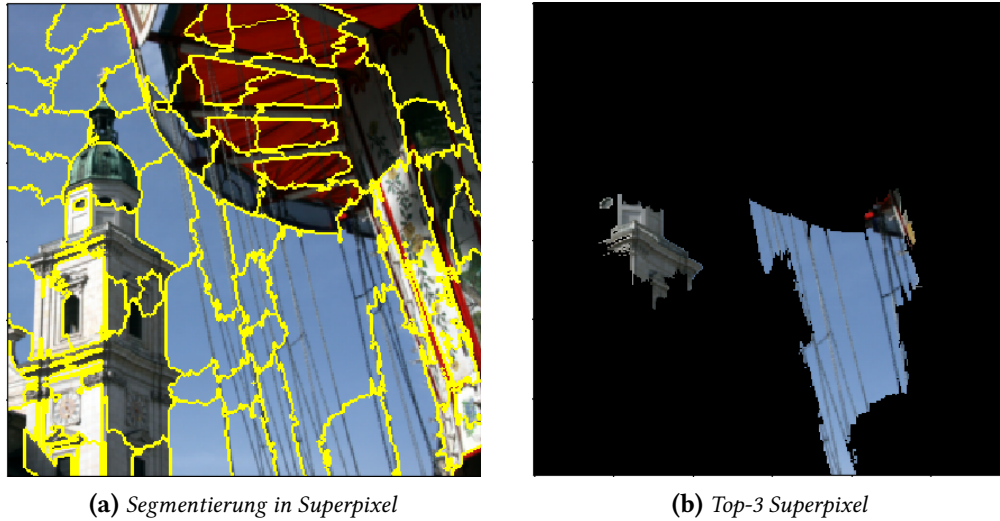


Abbildung 3.3: LIME-Erklärung: Zu sehen ist ein in Superpixel zerlegtes Foto (3.3a) und die drei einflussreichsten Superpixel (3.3b) für dessen Klassifikation als ‚Hängebrücke‘ durch das Modell Inception V3 (0,07%). Womöglich hielt das Modell die Ketten des Karussells für Tragseile. Bildquellen: Privat / eigene Darstellung unter Verwendung von LIME (Ribeiro et al., 2016). Siehe Teil IV für Details.

Der Erklärbarkeits-Leistungsfähigkeits-Tradeoff

Ein immer wieder vorgebrachter Einwand gegen ante-hoc erklärbare Systeme ist ihre geringere Leistungsfähigkeit. XAI soll einen Ausweg aus dem resultierenden Dilemma bieten, indem leistungsfähige, aber opake KI-Systeme durch Post-hoc-Methoden verständlich gemacht werden. Der vielzitierte ‚Erklärbarkeits-Performance-Tradeoff‘, demzufolge Leistungsfähigkeit und Opazität eines Systems proportional zueinander sind, ist in dieser Pauschalität allerdings nicht haltbar (Crook et al., 2023; Rudin, 2019). In vielen Anwendungsbereichen (eine wichtige Ausnahme ist die Verarbeitung von Audio- oder Bilddaten) kann die gewünschte Leistungsfähigkeit ebenso gut durch ein ante-hoc erklärbares System erzielt werden, ohne auf aufwändige und womöglich fehleranfällige Post-hoc-Methoden angewiesen zu sein.

Dennoch dominieren (systemagnostische, siehe unten) Post-hoc-Ansätze gegenüber ante-hoc erklärbaren Modellen die derzeitige XAI-Forschung. Neben dem Erklärbarkeits-Leistungsfähigkeits-Tradeoff steckt dahinter der Wunsch, leistungsfähige KI-Modelle für vielfältige Anwendungsbereiche zu entwickeln, ohne den vergleichsweise aufwändigen Entwicklungsprozess ante-hoc erklärbarer KI auf sich nehmen zu müssen (Sokol & Flach, 2022, S. 15).²⁵ Doch derzeitige Post-hoc-Methoden liefern nicht immer jene Erklärungen,

²⁵Das zeigt, dass in den Zielkonflikt zwischen Erklärbarkeit und Leistungsfähigkeit zusätzliche Faktoren hineinspielen, etwa die Verfügbarkeit von Ressourcen (z. B. Zeit, Expertise oder Geld) während der Ent-

die sich gemeinhin erhofft werden: Viele Methoden sind anfällig für Manipulationen durch böswillige Akteure (Bordt et al., 2022; Slack et al., 2020). Experimente legen nahe, dass manche Ansätze weitestgehend unabhängig von den Parametern des zu erklärenden KI-Modells und den Trainingsdaten sind und dass einige Methoden nicht die Sensitivität des Modells gegenüber Veränderungen der Eingaben widerspiegeln (eine Übersicht bietet Molnar, 2022, Kap. 9.2.8). Solche Befunde wecken Zweifel an der Akkuratheit der generierten Erklärungen. Hinzu kommt, dass Post-hoc-Ansätze unterschiedliche Herangehensweisen wählen, um die Ausgaben von KI-Systemen zu erklären. Diese Unterschiede führen dazu, dass sich für dieselbe Ausgabe desselben Modells radikal verschiedene Erklärungen generieren lassen, indem man mehrere XAI-Ansätze verwendet – ohne dass die ‚beste‘ oder ‚wahre‘ Erklärung bestimmt werden könnte (Bordt et al., 2022, Kap. 4.3).

Darin besteht ein fundamentales Problem von Post-hoc-Ansätzen: Ihre Akkuratheit kann in der Regel nur anhand von Experimenten abgeschätzt, aber nicht ohne Zweifel festgestellt werden (ebd., S. 899; Samek et al., 2021, S. 253). Denn die exakte Funktionsweise von hochgradig opaken Systemen wie neuronalen Netzen ist (auf Ebene menschenverständlicher Repräsentationen) unbekannt – schließlich gibt es daher überhaupt erst Post-hoc-Methoden, die versuchen, das zu ändern. Wir stehen also vor dem „epistemischen Problem zweiter Ordnung“ (Grote et al., 2024, S. 7; meine Übersetzung), die Zuverlässigkeit von XAI-Methoden zu evaluieren, um auf dieser Basis die Zuverlässigkeit des zugrundeliegenden KI-Systems bewerten zu können. Angesichts dessen müssen XAI-Forscher:innen sich die Frage gefallen lassen, ob die Entwicklung erklärbarer KI-Systeme mithilfe von Post-hoc-Methoden eigentlich nicht ebenso aufwändig ist wie das Entwickeln ante-hoc erklärbarer Systeme (vgl. Rudin, 2019; Sokol & Flach, 2022, S. 21).

3.2.2 Geltungsbereich: global versus lokal

Die Differenzierung zwischen globaler und lokaler Erklärbarkeit betrifft den Geltungsbereich der Erklärung, die von einem Erklärbarkeitsansatz generiert wird. Lokale Erklärungen sind für eine einzelne Eingabe (oder eine kleine Teilmenge der Eingaben) sowie in deren Nachbarschaft gültig; globale Erklärungen besitzen Gültigkeit in Bezug auf alle (validen) Eingaben des Modells (Schwalbe & Finzel, 2024, S. 3075). Wenngleich unüblich, ist auch eine Kategorie zwischen ‚lokal‘ und ‚global‘ hilfreich (nennen wir sie ‚glokal‘) – beispielsweise, wenn man eine Erklärung dafür haben möchte, warum ein Bildklassifikations-System bestimmte Bilder als ‚Stoppschild‘ klassifiziert: Weder wird das gesamte Modell erklärt, noch eine einzelne Ausgabe, sondern eine größere Gruppe von Ausgaben (siehe dazu Sokol & Flach, 2020, S. 2).

wicklungsphase (Crook et al., 2023).

Globale und globale Erklärbarkeit

Die im vorherigen Abschnitt vorgestellten ante-hoc erklärbaren Systeme können als Variante globaler Erklärbarkeit aufgefasst werden. Darüber hinaus gibt es auch verschiedene globale Post-hoc-Methoden. Ein einfaches Beispiel sind globale Surrogat-Modelle. Dafür trainiert man ein ante-hoc erklärbares Modell (z. B. einen Entscheidungsbaum) mit den Vorhersagen eines architektonisch opaken Systems (etwa eine Stützvektormaschine oder ein neuronales Netz). Das Surrogat approximiert das Verhalten des ursprünglichen Systems, ist im Gegensatz dazu aber (besser) verständlich (Molnar, 2022, Kap. 7.6).

Viele andere als global eingestufte Ansätze sind streng genommen eher global. Ein solcher Ansatz, der auf bildverarbeitende CNN-Modelle gerichtet ist, heißt Feature-Visualisierung (Olah et al., 2017). Die Methode generiert künstliche Bilder, die einen spezifischen Teil eines CNN stark aktivieren. Anhand dessen lässt sich abschätzen, auf welche Features der jeweils betrachtete Teil des Netzwerks reagiert. Dadurch lassen sich zum Beispiel typische Merkmale von Mitgliedern einer ganzen Bildklasse visualisieren. Ein ähnlicher Ansatz wählt natürliche Bilder aus den Trainingsdaten aus, die bestimmte Neuronen maximal aktivieren (*dataset examples*; Schubert et al., 2021). Durch das Auffinden ‚einflussreicher Beispiele‘ (engl. *influential instances*) lassen sich hingegen Trainingsdaten ausfindig machen, die während der Lernphase großen Einfluss auf die Modell-Parameter hatten (Koh & Liang, 2018; Molnar, 2022, Kap. 9.5). Und die SpRAY-Methode (*Spectral Relevance Analysis*) kombiniert mehrere lokale Erklärungen und identifiziert dadurch globale Entscheidungsmuster (Lapuschkin et al., 2019).

Globale bzw. globale Methoden sind hilfreich, wenn man das durchschnittliche Verhalten eines KI-Systems verstehen möchte, etwa um systematische Verzerrungen und Fehler aufzudecken oder um nachzuvollziehen, wie ein KI-System grundsätzlich funktioniert (Molnar, 2022, S. 115). Die Mehrheit der jüngeren XAI-Ansätze, darunter auch die meisten Attributionsmethoden (Abschnitt 3.2.1), bemüht sich hingegen um lokale Erklärbarkeit. Nicht nur sind lokale Erklärungen einfacher zu erzielen, sie sind auch für bestimmte Anwendungsfälle relevanter. Ein Verständnis aller oder der übergeordneten Strategien, die ein System verwendet, ist oft wenig hilfreich, wenn man an den Gründen für eine bestimmte Entscheidung interessiert ist.

Lokale Erklärbarkeit

Für die lokale Erklärung von Bildklassifikationssystemen werden häufig die oben erwähnten Salienzkarten verwendet. Diese Salienzkarten heben jene Bildbereiche hervor, die laut der verwendeten Erklärbarkeitsmethode den größten Einfluss auf die Klassifikation des fraglichen Bildes hatten. Ähnliche Erklärungen können auch für andere Datenformate, wie Text- oder tabellarische Daten, generiert werden. Beispielsweise kann LIME durch Aufzeigen einflussreicher Wörter erklären, weshalb ein Text einem bestimmten Thema zugeordnet wurde (Ribeiro et al., 2016, Kap. 3.5). Insgesamt dominieren allerdings

XAI-Methoden für Systeme, die Bilddaten verarbeiten (Nauta et al., 2023, Abb. 5 auf S. 14).

Eine andere Gruppe lokaler Erklärbarkeitsansätze sind so genannte kontrafaktische Erklärungen. Solche Erklärungen geben die kleinstmögliche Veränderung der Features der Eingabedaten an, die den Output des Systems von einem unerwünschten in eine bestimmte, erwünschte Ausgabe umschlagen lässt (Molnar, 2022, S. 194; Wachter et al., 2018, S. 844). Kontrafaktische Erklärungen sollen vor allem das Handeln von Personen derart anleiten, dass diese ein bestimmtes Ergebnis erzielen können. Eine kontrafaktische Erklärung für ein Kreditvergabesystem könnte beispielsweise angeben, wie hoch das Gehalt einer abgelehnten Antragstellerin hätte sein müssen, um einen Kredit bewilligt zu bekommen.

Ein wichtiges Kriterium für gute kontrafaktische Erklärungen ist, dass die daraus abgeleiteten Handlungsempfehlungen (z. B. das eigene Einkommen zu steigern) für das entsprechende Individuum realistischerweise umgesetzt werden können (Karimi et al., 2021; Wachter et al., 2018). Manche kontrafaktischen Erklärungen enthalten für niemanden eine realistische Handlungsempfehlung (etwa, wenn man das eigene Alter reduzieren müsste, um das gewünschte Ergebnis zu erreichen). In anderen Fällen unterscheiden sich die Handlungsoptionen von Person zu Person (manche Personen können ihr Einkommen z. B. einfacher steigern als andere). Kontrafaktische Erklärungen sind daher nur dann hilfreich, wenn sie auf die individuellen Möglichkeiten der betreffenden Person abgestimmt sind.

Eine weitere Anforderung an kontrafaktische Erklärungen besteht darin, dass die abgeleiteten Handlungsempfehlungen auch tatsächlich eine sinnvolle und effektive Strategie darstellen, den gewünschten Ausgang herbeizuführen. Das ist leider nicht immer der Fall. Im Hintergrund kontrafaktischer Erklärungen steht die (häufig falsche) Annahme, dass die vom Modell verwendeten Variablen (z. B. Gehalt, Alter, Ersparnisse...) unabhängig voneinander sind. In der realen Welt außerhalb des Modells geht die Veränderung einer Variable aber mitunter mit Veränderungen in einer anderen Variable einher. Beispielsweise kann das Steigern des eigenen Einkommens eine Weile dauern, dadurch wird sich auch das Alter der Antragstellerin verändern. Und ein höheres Einkommen lässt häufig auch die Ersparnisse anwachsen. Da solche kausalen Zusammenhänge von KI-Systemen oft nicht berücksichtigt werden, geben kontrafaktische Erklärungen mitunter unrealistische oder suboptimale Handlungsempfehlungen ab (Karimi et al., 2021).

3.2.3 Anwendbarkeit: systemagnostisch versus systemspezifisch

Schließlich lassen sich XAI-Methoden auch hinsichtlich der Breite ihrer Anwendbarkeit unterscheiden. Systemagnostische Methoden sind auf jeden Typ von KI-Systemen anwendbar, systemspezifische Methoden funktionieren nur für einen bestimmten Systemtyp (z. B. neuronale Netzwerke oder Stützvektormaschinen). Normalerweise wird diese Unterscheidung lediglich auf Post-hoc-Methoden angewandt (Speith, 2022, S. 2240). Die Bezeichnung ‚systemagnostisch‘ ist daher implizit oder explizit häufig für Methoden reserviert, die auf jede Art von *ML-basierten* KI-Systemen angewandt werden können (vgl. etwa Guidotti

et al., 2019, S. 93:17; Molnar, 2022, S. 109). In Anbetracht dessen, dass auch konventionelle Systeme mitunter nur durch zusätzliche Erklärungsstrategien wie Visualisierungen oder Erklärungsmodule verständlich gemacht werden können (siehe Abschnitte 2.2.1 und 3.2.1), scheint diese Einschränkung aber überflüssig zu sein.

Die breite Anwendbarkeit systemagnostischer Methoden kommt daher, dass sie die spezifischen Eigenschaften des zu erklärenden Systems, wie z. B. seine Modellarchitektur, nicht berücksichtigen (Guidotti et al., 2019, S. 93:19). Stattdessen analysieren viele dieser Methoden lediglich die Eingaben und zugehörigen Ausgaben des Systems und verknüpfen beides durch ein ante-hoc erklärbares Modell (ebd.; Schwalbe & Finzel, 2024, S. 3072). Zwar sind systemagnostische Methoden aufgrund ihrer Vielseitigkeit attraktiv. Allerdings ist fraglich, ob damit generierte Erklärungen akkurat abbilden, was innerhalb des Systems passiert – schließlich fließt die spezifische Funktionsweise des erklärten Systems häufig nicht direkt in den Erklärbarkeitsansatz ein (Rudin, 2019, S. 207–208).

3.2.4 Evaluationskriterien für XAI-Erklärungen

Angesichts der Schwächen einiger Erklärbarkeitsmethoden stellt sich die Frage, wie XAI-Erklärungen evaluiert werden können. Anfangs herrschte in der informatischen Erklärbarkeitsforschung eine recht hemdsärmelige Herangehensweise an diese Frage vor: Eine gute Erklärung zeichnet sich demnach vor allem dadurch aus, für die menschlichen Adressat:innen ‚intuitiv plausibel‘ zu sein, d. h. mit (bewussten) menschlichen Denkprozessen übereinzustimmen (vgl. Freiesleben & König, 2023, S. 53–54; Molnar, 2022, S. 262).

Dieses Evaluationskriterium birgt zwei Probleme: Erstens ist die Übereinstimmung mit menschlichen Intuitionen kein Gradmesser der Akkuratheit bzw. Wahrheitstreue einer Erklärung (engl. *faithfulness*). Die Funktionsweise von KI-Systemen weicht in vielerlei Hinsicht vom menschlichen Denken ab, entsprechend kann man nicht erwarten, dass Erklärungen, die das Verhalten des Systems wahrheitsgetreu wiedergeben, für menschliche Betrachter:innen stets plausibel wirken (Selbst & Barocas, 2018). Umgekehrt kann eine intuitiv überzeugende Erklärung das Ergebnis einer irreführenden XAI-Methode sein. Zweitens droht der Bestätigungsfehler, die Interpretation einer Erklärung fehlzuleiten (Freiesleben, 2026, S. 286; Ghassemi et al., 2021, S. 746). Der Bestätigungsfehler ist eine kognitive Verzerrung und beschreibt die menschliche Tendenz, Informationen zu bevorzugen, die im Einklang mit den eigenen Überzeugungen stehen. Uneindeutige Erklärungen können daher so interpretiert werden, dass sie zu den bestehenden Annahmen der Betrachterin passen – und nicht so, dass wahre Aussagen daraus gewonnen werden.²⁶

Es stellt sich also die Frage danach, woran man die Qualität einer XAI-Erklärung erkennen kann. In der Wissenschaftstheorie wird diese Frage traditionell unter dem Schlagwort ‚explanatorische Tugenden‘ verhandelt. Die wohl bekanntesten explanatorischen Tugenden hat Thomas Kuhn formuliert: Erklärungen sollten möglichst (1) einfach sowie (2) empirisch

²⁶Auf die problematische Vagheit visueller XAI-Erklärungen komme ich in Abschnitt 7.3.2 erneut zurück.

akkurat sein, (3) einen großen Geltungsbereich besitzen, (4) in sich konsistent sein und nicht im Widerspruch zu anderen akzeptierten Theorien oder Erklärungen stehen sowie (5) sich als fruchtbar für weitere Forschung erweisen (Kuhn, 1977, S. 321–322). Streng genommen handelt es sich bei diesen fünf Merkmalen guter Erklärungen um *theoretische* Tugenden – also um wünschenswerte Eigenschaften wissenschaftlicher Theorien, anhand derer Wissenschaftler:innen zwischen konkurrierenden Theorien wählen können. Nichtsdestotrotz werden diese Tugenden auch häufig für die Charakterisierung guter Erklärungen herangezogen.

Kuhns Tugenden lassen sich auch auf XAI anwenden. Eine gute XAI-Erklärung zeichnet sich demnach dadurch aus, dass sie möglichst einfach ist und mit dem übereinstimmt, was wir bereits über das KI-System wissen. Sie sollte bestenfalls nicht nur für eine einzelne Ausgabe gültig sein, sondern ein größeres Spektrum an Ausgaben erklären können. (Allerdings wurde in Abschnitt 3.2.2 deutlich, dass im Kontext von XAI lokale Erklärungen mitunter hilfreicher sind als globale Erklärungen.) Natürlich sollte eine XAI-Erklärung frei von Widersprüchen sein und idealerweise auch einen Ausgangspunkt für weitere Nachforschungen liefern.

Auch innerhalb der informatischen XAI-Forschung haben sich in jüngerer Zeit verschiedene Autor:innen darum bemüht, aussagekräftige Qualitätsstandards für XAI-Erklärungen zu etablieren – viele davon ähneln Kuhns explanatorischen Tugenden. Beispielsweise bewerten Wojciech Samek et al. unterschiedliche Post-hoc-Erklärungen anhand deren Übereinstimmung mit dem zugrundeliegenden Modell,²⁷ der Verständlichkeit für Menschen und der praktischen Anwendbarkeit der Methode (Samek et al., 2021, Kap. IV). Aniek Markus et al. schlagen vier Eigenschaften von Erklärbarkeit vor, anhand derer sich XAI-Erklärungen evaluieren lassen: Erklärungen sollten (1) klar bzw. nicht ambig sowie möglichst (2) kompakt und sparsam sein sowie die Ausgabe möglichst (3) vollständig und (4) wahrheitsgemäß erklären (Markus et al., 2021, S. 2–3). Jianlong Zhou et al. ergänzen noch das Kriterium der (5) Breite: Eine gute Erklärung sollte eine möglichst breite Gültigkeit besitzen (J. Zhou et al., 2021, S. 3).

Einen besonders ausführlich ausgearbeiteten Vorschlag machen Meike Nauta et al. Die Autor:innen schlagen eine Taxonomie von zwölf Eigenschaften vor, anhand derer sich die Qualität von XAI-Erklärungen evaluieren lässt (Nauta et al., 2023). Neben Eigenschaften, die auch in der Wissenschaftstheorie als explanatorische Tugenden gelten (z. B. Korrektheit, Vollständigkeit, Konsistenz, Kompaktheit) identifizieren Nauta et al. noch diverse weitere wünschenswerte Eigenschaften von XAI-Erklärungen. Einige davon sollen sicherstellen, dass die Erklärung den spezifischen Anforderungen des Kontexts entspricht: Erklärungen sollten zur Fragestellung und dem Vorwissen der Adressat:innen passen, kontrastiv und interaktiv sein sowie ein geeignetes Format besitzen (ebd., Kap. 3.3).

Auch wenn die XAI-Forschung sich zuletzt vermehrt mit Qualitätskriterien für Erklä-

²⁷Wie in Abschnitt 3.2.1 dargelegt, ist diese Eigenschaft bei vielen architektonisch opaken Systemen (z. B. neuronalen Netzen) in Ermangelung der ‚ground truth‘ nur schwer zu evaluieren.

rungen befasst hat: Die Frage nach Eigenschaften guter Erklärungen für KI-Systeme reicht mindestens in die Zeit der Expertensysteme zurück. Beispielsweise identifizierten William Swartout und Johanna Moore bereits zu Beginn der 1990er Jahre diverse Desiderata für Erklärungen solcher Systeme. Erklärungen sollten unter anderem wahrheitsgetreu wiedergeben, was im System passiert, für Entwickler:innen einfach zu implementieren sein und die Leistungsfähigkeit des Systems nicht einschränken (Swartout & Moore, 1993, S. 545, 547).

Weitaus interessanter sind aber die Desiderata der Verständlichkeit und Zulänglichkeit (engl. *sufficiency*). Um *verständlich* zu sein, sollten Erklärungen laut Swartout & Moore zum Beispiel zu dem Hintergrundwissen, den Zielen und den Präferenzen der Adressat:innen passen, lediglich bekannte Terminologie verwenden bzw. unbekannte Ausdrücke erläutern und interaktiv sowie adaptiv sein. Letzteres beinhaltet auch, dass Erklärungen in unterschiedlichen Detail- und Abstraktionsgraden und unter verschiedenen Gesichtspunkten formuliert werden können. *Zulänglichkeit* erfordert, dass das System eine hinreichende Menge an Wissen repräsentiert, um fundierte Erklärungen abgeben und Nutzerfragen beantworten zu können. Die Erklärung des Verhaltens des Systems (also das, womit sich XAI heute vornehmlich befasst) macht davon nur einen kleinen Teil aus. Zusätzlich sollten Expertensysteme laut den Autor:innen auch ihr Verhalten begründen können und zum Beispiel erläutern, warum das System in einer bestimmten Form gestaltet worden ist, Hintergrundinformationen über die Anwendungsdomäne liefern und Fachbegriffe erklären (ebd., S. 545–547). Im Vergleich dazu scheinen die heutigen Ziele von XAI weitaus bescheidener. Dennoch sind wir derzeit weit davon entfernt, für alle Kontexte Erklärungen zu generieren, die relevanten Qualitätsstandards gerecht werden.

3.3 KI-Systeme erklären oder verstehen?

Der informatische Forschungszweig der erklärbaren KI – *nomen est omen* – zielt darauf ab, KI-Systeme oder einzelne ihrer Aspekte zu erklären und dadurch für Menschen verständlich zu machen, und XAI-Forscher:innen entwickeln diverse Erklärbarkeitsmethoden, deren Ausgaben sie selbstverständlich als ‚Erklärungen‘ bezeichnen. Die Natur der Erklärung, explanatorische Tugenden oder das Verhältnis von Erklärungen und Verstehen sind wiederum traditionelle Themen der Wissenschafts- und Erkenntnistheorie. Angesichts dessen ist es naheliegend, philosophische Überlegungen zu Erklärungen für das Erklären von KI-Systemen fruchtbar zu machen.

In diesem Abschnitt möchte ich einige Schwierigkeiten aufzeigen, die sich bei diesem Unterfangen stellen. In Abschnitt 3.3.1 erläutere ich, weshalb sich viele philosophische Theorien der Erklärung nicht unmittelbar auf KI-Systeme anwenden lassen. Daher ist uneindeutig, ob KI-Systeme im Sinne bestehender Theorien als erklärbar gelten können. Und selbst wenn KI-Systeme im philosophischen Sinne erklärbar sein sollten, sind Erklärungen nicht der einzige Weg, um relevante Aspekte dieser Systeme verständlich zu

machen (Abschnitt 3.3.2). Ich schlage daher vor, Verstehensermöglicher in XAI-Kontexten als breitere Oberkategorie aufzufassen, die Erklärungen einschließt, aber darüber hinaus auch andere Repräsentationen umfasst.

3.3.1 Sind KI-Systeme erklärbar?

Die Philosophie hat diverse Vorschläge dazu hervorgebracht, was eine Erklärung ist. Vereinfacht gesprochen, handelt es sich bei Erklärungen um Antworten auf Warum-Fragen (diese Charakterisierung findet sich bereits in der antiken Philosophie, siehe z. B. Aristoteles, Phys. II 3 ca. 194b16–195a3). Seit den Arbeiten von Carl Gustav Hempel und Paul Oppenheim (1948) zu wissenschaftlichen Erklärungen wird häufig zwischen zwei Erklärungskomponenten unterschieden. Das *Explanans* bezeichnet den erklärenden Teil der Erklärung, das *Explanandum* ist dasjenige, das erklärt werden soll. Ein Beispiel: Willis Haare haben einen Grünstich (*Explanandum*), weil er mit blondiertem Haar im Pool war und Bestandteile des Wassers mit seinen Haaren reagiert haben (*Explanans*). Jenseits solch basaler Annahmen existiert allerdings kein philosophischer Konsens, sondern eine Reihe konkurrierender Erklärungstheorien.

Viele Kontroversen hängen mit der Frage danach zusammen, ob ein ontischer oder ein epistemischer bzw. kommunikativer Ansatz die Natur der Erklärung am besten erfasst. Diese Ontisch-epistemisch-Debatte (z. B. Craver, 2014; Illari, 2013; Kästner & Haueis, 2021; Potochnik, 2016) ist umfangreich und komplex.²⁸ An dieser Stelle genügt die folgende Charakterisierung der Unterscheidung (Kästner & Haueis, 2021, S. 1636): *Ontische Theorien* nehmen an, dass Erklärungen Dinge in der Welt sind – beispielsweise Ursachen oder Mechanismen, die ein Phänomen hervorbringen. Diesem Paradigma zufolge sollte eine Erklärung möglichst akkurat und vollständig sein. *Epistemische Theorien* betrachten Erklärungen hingegen als (z. B. linguistische) Repräsentationen. Hier steht der kommunikative Akt des Erklärens oft stärker im Vordergrund, bzw. der epistemische Zustand, der dadurch hervorgebracht werden soll. Epistemischen Ansätzen zufolge sollte eine Erklärung das *Explanandum* daher möglichst verständlich oder nachvollziehbar machen (ebd.). In der Regel erkennen beide Seiten die Relevanz der jeweils anderen Perspektive zumindest implizit an (vgl. Fleisher, 2022, Fn. 10 auf S. 11; Illari, 2013; Kästner & Haueis, 2021; Potochnik, 2016). Strittig ist vielmehr, wie viel Gewicht das ontische bzw. epistemische Element in einer Theorie der Erklärung bekommen sollte.

²⁸Ein wichtiger Grund dafür ist, dass die Unterscheidung, je nach Autor:in, diverse Fragestellungen einbezieht – etwa die Frage nach dem Einfluss kontextueller bzw. pragmatischer Faktoren, die Realismus-Antirealismus-Debatte (vgl. Fußnote 79), explanatorische Tugenden (Abschnitt 3.2.4) oder das epistemische Ziel von Erklärungen. Weiterhin lassen sich einzelne Erklärungstheorien oder ‚Theoriefamilien‘ nicht immer eindeutig dem einen oder anderen Lager zuordnen. Beispielsweise vertreten Autor:innen wie Carl Craver (z. B. Craver, 2014; Craver & Darden, 2013) eine ontische Theorie der mechanistischen Erklärung, während Autor:innen wie William Bechtel (z. B. Bechtel, 2011; Bechtel & Abrahamsen, 2005) für einen epistemischen Ansatz der mechanistischen Erklärung plädieren; und Carl Hempels D-N-Modell weist sowohl ontische als auch epistemische Elemente auf (Craver, 2014, S. 31).

Für meine Zwecke ist es hilfreich, eine epistemische Perspektive einzunehmen und Erklärungen als Repräsentationen aufzufassen, die darauf abzielen, Verstehen hervorzu-
bringen. Die Inhalte solcher Repräsentationen sind explanatorische Informationen – also
Informationen über Abhängigkeitsbeziehungen zwischen Explanans- und Explanandum-
Phänomen (vgl. J. Kim, 1994, S. 68; McCain, 2016, S. 143–144). Eine Erklärung repräsentiert
solche Abhängigkeitsbeziehungen und gibt dadurch bestenfalls Aufschluss über Aspekte
der Welt – hier kommt also auch die ontische Perspektive auf Erklärungen ins Spiel. Die
Redeweise von Erklärungen als Repräsentationen fügt sich gut in die XAI-Literatur ein,
wo beispielsweise Sätze in natürlicher Sprache oder Salienzkarten als Erklärungen betrach-
tet werden (wenngleich viele Philosoph:innen nicht-linguistische Repräsentationen als
Erklärungen ausschließen – ich komme in Abschnitt 3.3.2 darauf zurück).

Eine weitere kontroverse Frage betrifft die Art der Abhängigkeitsbeziehungen, die Er-
klärungen konstituieren bzw. repräsentieren. Wie nachfolgend deutlich werden wird, hat
die Philosophie der Erklärung darauf eine Reihe unterschiedlicher Antworten formuliert.
Bei der Mehrheit dieser Antworten zeigt sich eine Diskrepanz zu dem Vorhaben von XAI:
Philosophische Erklärungstheorien streben üblicherweise danach, wissenschaftliche Erklä-
rungen natürlicher Phänomene zu charakterisieren, manche betrachten auch Erklärungen
menschlicher Handlungen. KI-Systeme bzw. einzelne ihrer Aspekte sind hingegen keine
natürlichen Phänomene oder Akteur:innen, sondern mathematische Artefakte. Im Kontext
erklärbarer KI haben wir es also mit grundlegend anderen Explananda zu tun.

Im verbleibenden Abschnitt möchte ich einige Schwierigkeiten aufzeigen, die sich daraus
für das Erklären von KI-Systemen ergeben. Ich diskutiere nachfolgend eine Auswahl
wichtiger Theorien der Erklärung, die bereits auf KI-Systeme übertragen worden sind. Mein
Ergebnis wird lauten, dass viele philosophische, insbesondere wissenschaftstheoretische
Theorien der Erklärung nur eingeschränkt auf KI-Systeme anwendbar sind. Daher ist
uneindeutig, ob KI-Systeme im philosophischen Sinne erklärbar sind.²⁹ Abschließend
werde ich einen Weg aufzeigen, wie mit dieser Schwierigkeit umgegangen werden kann.

Das D-N-Modell

Die wohl bekannteste Erklärungstheorie ist das deduktiv-nomologische (D-N) Modell der
wissenschaftlichen Erklärung von Carl Hempel (Hempel & Oppenheim, 1948; Hempel,
1965), demzufolge eine Erklärung die Struktur eines schlüssigen deduktiven Arguments
besitzt. Die (wahren) Prämissen bilden das Explanans und bestehen aus mindestens einem
allgemeinen Naturgesetz und typischerweise auch aus Antezedensbedingungen, die den

²⁹Um diese Frage abschließend zu beantworten, bräuchte es eine ausführlichere – und vor allem alle ver-
fügbaren Vorschläge abdeckende – Diskussion von Erklärungstheorien. Mein Ziel ist bescheidener: Ich
möchte einige Diskrepanzen zwischen den Erfordernissen in XAI-Kontexten und den in der Philosophie
gängigerweise verhandelten Erklärungen aufzeigen. Diese Unterschiede sprechen in meinen Augen dafür,
dass wir eine auf die Besonderheiten von KI abgestimmte Erklärungstheorie und/oder eine Theorie nicht-
explanatorischer Verstehensermöglicher brauchen, um XAI-Erklärungen philosophisch zu untermauern.

konkreten Fall beschreiben. Das Explanandum ist die logische Konklusion aus Naturgesetzen und Antezedensbedingungen. Beispielsweise kann man mit dem D-N-Schema erklären, warum ein Stift zu Boden fällt, wenn man ihn loslässt. Das Gravitationsgesetz fungiert als Naturgesetz und die Antezedensbedingungen beschreiben die konkrete Situation (ein Stift, auf den neben der Gravitation keine weiteren Kräfte wirken usw.). Dass der Stift zu Boden fällt, lässt sich dann aus dem Explanans deduzieren.

Adrian Erasmus et al. haben Theorien der wissenschaftlichen Erklärung auf das Erklären neuronaler Netze in medizinischen Kontexten übertragen, darunter auch das D-N-Modell. Die Autoren argumentieren, dass Hempels Theorie sinnvoll auf solche Systeme angewandt werden kann (Erasmus et al., 2021, S. 844): Die allgemeinen Gesetze des Explanans entsprechen vornehmlich den Parametern des Modells, gemäß derer die Ausgabe produziert wird. Die Antezedensbedingungen sind dann die spezifische Eingabe, die gemäß der ‚Gesetze‘ verarbeitet und in eine Ausgabe (das Explanandum) umgewandelt wird.

Allerdings ist zweifelhaft, ob die Parameter eines neuronalen Netzwerks allgemein genug sind, um als Gesetzmäßigkeiten im Sinne Hempels durchzugehen. Wie Andrés Páez feststellt, handelt es sich bei diesen Parametern um sehr spezifische Anpassungen an konkrete Trainingsdaten, nicht um generelle Gesetzmäßigkeiten (Páez, 2026, S. 140). Wenn Páez Recht hat, sind zumindest neuronale Netze im Sinne des D-N-Modells daher vermutlich nicht erklärbar. Doch selbst wenn dem so wäre: Es ist fraglich, ob eine solche Erklärung irgendeinen Nutzen mit sich bringt. Wenn die D-N-Erklärung eines KI-Systems mehrere Millionen ‚Gesetzmäßigkeiten‘ in Form numerischer Parameter aufruft, wird sie kaum zur Verständlichkeit des Systems beitragen.

Erasmus et al. grenzen (objektive) Erklärbarkeit daher auch ausdrücklich von (teils subjektiver) Interpretierbarkeit bzw. Verständlichkeit ab (Erasmus et al., 2021, Kap. 3.2). Unverständliche Erklärungen können den Autoren zufolge zum Interpretandum einer Interpretation werden, die bestenfalls eine verständliche Erklärung hervorbringt. Eine solche Interpretation kann beispielsweise erfolgen, indem man eine globale auf eine lokale Erklärung herunterbricht (siehe Abschnitt 3.2.2), und/oder indem man die komplexen nicht-linearen Funktionen eines neuronalen Netzwerks linear approximiert (ebd., Kap. 4; diesen Ansatz wählt z. B. die LIME-Methode). Die eigentliche Herausforderung von Erklärbarkeit, nämlich KI-Systeme bzw. einzelne ihrer Aspekte verständlich zu machen, wird also auf den Schritt der Interpretation verlagert.

Kausal-mechanische Erklärungen

Kausalen Erklärungstheorien zufolge gibt eine Erklärung die Ursachen für ein Ereignis an. Beispielsweise besteht die Grundidee von Wesley Salmon (1984) kausal-mechanischer Theorie der Erklärung darin, bestimmte kausale Prozesse und Interaktionen zu benennen, die aufzeigen, wie sich ein Phänomen in die kausale Struktur der Welt einfügt. Ein kausaler Prozess zeichnet sich durch Kontinuität in Raum und Zeit aus. Ein *echter* kausaler Prozess unterscheidet sich von so genannten Pseudo-Prozessen dadurch, dass er eine bestimmte

physikalische Struktur während räumlich-zeitlicher Veränderungen ohne zusätzliche Eingriffe von außen aufrechterhalten kann – er kann ein Merkmal weitergeben (engl. *mark transmission*). Eine Kugel, die sich über einen Billardtisch bewegt, trägt beispielsweise einen Bewegungsimpuls weiter – die Bewegung der Kugel stellt einen echten kausalen Prozess dar. Ein Schatten ist hingegen ein Pseudo-Prozess. Er kann selbst kein Merkmal weitergeben; seine Form und Intensität hängt gänzlich von anderen Faktoren ab (z. B. der Lichtquelle oder etwaigen Hindernissen). Kausale Interaktionen finden statt, wenn zwei kausale Prozesse interagieren, z. B. wenn zwei Kugeln aufeinanderprallen und ihre Bewegungsrichtung verändern.

Erasmus et al. argumentieren, dass auch Salmons Theorie auf opake KI-Systeme angewandt werden kann: Die relevanten kausalen Prozesse sind das Einspeisen der Eingabedaten in das System und deren Verarbeitung innerhalb der verschiedenen Schichten des Netzwerks. Die ‚Signale‘, die zwischen Schichten und Neuronen wandern, entsprechen dann Salmons kausalen Interaktionen (Erasmus et al., 2021, S. 845). Durch eine solche kausal-mechanische Erklärung könnte man dann beispielsweise einzelne Ausgaben eines neuronalen Netzes erklären.

Doch bei genauerer Betrachtung ergeben sich auch bei der Anwendung des kausal-mechanischen Modells auf KI-Systeme verschiedene Schwierigkeiten. Ein Problem besteht darin, dass die Redeweise von Schichten und Signalen in neuronalen Netzwerken lediglich metaphorisch ist (siehe Abschnitt 2.1.2). Neuronale Netze sind mathematische Strukturen, deren Elemente nicht in kausalen Zusammenhängen stehen können (Páez, 2026, S. 143); das Gleiche gilt auch für die z. B. logischen Operationen konventioneller KI. Daher scheidet eine genuine kausale Erklärung aus, wenn man KI-Systeme auf abstrakt-mathematischer Ebene betrachtet.

Alternativ könnte man KI-Systeme auch auf Ebene der konkreten physischen Realisierung der mathematischen Operationen in einer bestimmten Hardware-Konfiguration erklären. Zwar kommen dann genuine kausale Zusammenhänge ins Spiel, doch zeigt sich ein anderes Problem: Wie Páez anmerkt, wird das Erklären von KI-Systemen auf dieser Betrachtungsebene bloß noch herausfordernder. Im Falle eines neuronalen Netzes müsste man dafür beispielsweise Korrelationen zwischen den mathematischen Operationen des Modells und den kausalen Hardware-Prozessen identifizieren (ebd.). Und selbst wenn dies gelänge, ist es in der Regel nicht hilfreich, die Ausgaben eines KI-Systems mithilfe von Schaltkreisen oder Binärzuständen zu erklären – die Verständlichkeit bliebe abermals auf der Strecke. Insgesamt betrachtet ist daher auch das kausal-mechanische Modell der Erklärung nur eingeschränkt auf KI-Systeme übertragbar.

Mechanistische Erklärungen

Theorien mechanistischer Erklärungen, vertreten unter anderem von William Bechtel und Adele Abrahamsen (Bechtel, 2011; Bechtel & Abrahamsen, 2005) sowie Peter Machamer, Carl Craver und Lindley Darden (Craver & Darden, 2013; Machamer et al., 2000), sind

vornehmlich in Biologie, Neurowissenschaft und anderen Lebenswissenschaften populär. Mechanistische Erklärungen geben Aufschluss über die funktionale Struktur eines (z. B. biologischen) Systems: Sie decken auf, wie sich ein bestimmtes Ergebnis oder Verhalten aus dem räumlich-zeitlichen Zusammenspiel von Prozessen oder Systemkomponenten ergibt. Manche mechanistischen Erklärungen sind ätiologisch und decken die kausalen Prozesse auf, die ein bestimmtes Produkt oder Verhalten verursachen. Lena Kästner und Lise Marie Andersen nennen das Beispiel der Proteinbiosynthese, also der Neubildung von Proteinen in Zellen: Es handelt sich um eine kausale Sequenz, an deren Ende ein ‚Produkt‘ (ein Protein) steht (Kästner & Andersen, 2018, S. 1). Andere mechanistische Erklärungen sind konstitutiv – das Ablaufen des Mechanismus konstituiert das zu erklärende Phänomen. Ein gängiges Beispiel ist das Aktionspotenzial, das die Reizweiterleitung zwischen Zellen (z. B. Neuronen) ermöglicht: Das Aktionspotenzial *besteht in* den relevanten Prozessen, die in den einzelnen Teilen einer solchen Zelle ablaufen (ebd., S. 2).

Lena Kästner und Barnaby Crook illustrieren am Beispiel von Forscher:innen rund um Christopher Olah, wie mechanistische Erklärungen neuronaler Netze erlangt werden können (Kästner & Crook, 2024). Olah et al. identifizierten eine funktionale Einheit eines bildverarbeitenden CNN: einen Kurven-Detektor. Das CNN hatte während des Trainings gelernt, geschwungene Linien in Bildern zu erkennen, unabhängig von deren Dicke, Farbe oder Ausrichtung. Mithilfe verschiedener Experimente und XAI-Methoden konnten die Forscher:innen eine Gruppe von Neuronen lokalisieren, die über mehrere Schichten des Netzwerks verteilt ist und diese spezifische Subfunktion realisiert (siehe z. B. Cammarata et al., 2020b). Idealerweise würde man sämtliche Subfunktionen des Netzwerks auf gleichem Wege lokalisieren und dadurch den gesamten Mechanismus der Bildverarbeitung in verständliche Elemente zerlegen. Eine solche Dekomposition in funktionale Einheiten und die Beschreibung ihrer Interaktionen wäre dann eine vollständige mechanistische Erklärung des neuronalen Netzwerks (vgl. auch Erasmus et al., 2021, S. 846).

Das Beispiel von Olah und Kolleg:innen zeigt, dass mechanistische Erklärungen für KI-Systeme fruchtbar gemacht werden können. Allerdings besteht hier dasselbe Problem wie bei der Anwendung von Salmons Erklärungstheorie auf KI: Mechanismen sind kausal. Das ist offensichtlich bei der ätiologischen Lesart mechanistischer Erklärungen; doch da die Interaktion zwischen den Komponenten eines Mechanismus üblicherweise ebenfalls als kausal aufgefasst wird, gilt dies auch für die konstitutive Lesart (Kästner & Andersen, 2018, S. 2). Sofern man KI-Systeme nicht auf der Hardware-Ebene erklären will, bleiben daher auch mechanistische Erklärungen von KI-Systemen bis zu einem gewissen Grad metaphorisch.

Manipulationistische Erklärungen

Eine weitere einflussreiche kausale Theorie der Erklärung stammt von James Woodward (2003). Sein manipulationistischer Ansatz beruht auf dem Prinzip der kontrafaktischen Abhängigkeit zwischen Explanans und Explanandum. (Andere wichtige kontrafaktische

Theorien stammen von David Lewis und Judea Pearl; Lewis, 1973; Pearl & Mackenzie, 2018.) Woodward zufolge ist (vereinfacht ausgedrückt) X eine Ursache von Y genau dann, wenn eine Veränderung von X zu einer Veränderung von Y führen würde, bzw. wenn die Veränderung von Y nicht erfolgt wäre, wäre X nicht verändert worden. Zentral ist der Begriff der *Intervention*: Woodward meint damit die gezielte, ideale Manipulation von X – ideal, insofern jede Veränderung von Y eindeutig auf die Veränderung von X zurückgeführt werden kann (dieses Prinzip kennt man aus randomisierten kontrollierten Studien, z. B. um die Wirksamkeit eines Medikaments zu überprüfen). Dadurch können Erklärungen Was-wäre-wenn-die-Dinge-anders-gewesen-wären-Fragen beantworten.

Den Grundgedanken der Intervention greifen auch die kontrafaktischen Erklärungen auf, die in der Erklärbarkeitsliteratur diskutiert werden (siehe Abschnitt 3.2.2). Um beispielsweise herauszufinden, was die Ursache für die Ablehnung einer Anwärtlerin auf einen Bankkredit war, kann man fragen, welche Aspekte der Eingabedaten hätten anders sein müssen, damit ihr ein Kredit gewährt worden wäre. Diese Faktoren können dann in einer Erklärung genannt werden und geben – im Idealfall – Aufschluss über den Zusammenhang zwischen Eingabe und Ausgabe.

Erneut lässt sich einwenden, dass KI-Systeme als mathematische Artefakte bestenfalls metaphorisch durch eine Woodward'sche Erklärung erklärt werden können. Allerdings hat Stefan Buijsman kürzlich einen Lösungsvorschlag präsentiert: Er stimmt mit Pérez darin überein, dass ML-basierte KI-Systeme als nicht-kausale, mathematische Objekte aufgefasst werden sollten, und dass die Anwendung von kausalen bzw. wissenschaftlichen Erklärungstheorien nicht zielführend sei. Buijsman schlägt stattdessen vor, XAI-Erklärungen anhand von Theorien der mathematischen Erklärung auszubuchstabieren, die Woodwards manipulationistischen Ansatz zu einer Theorie der nicht-kausalen Erklärung umarbeiten (Buijsman, 2026). Wenn dies gelingt, wären zumindest ML-basierte KI-Systeme womöglich doch genuin erklärbar.

Doch die Anwendung manipulationistischer Erklärungen auf KI-Systeme wird auch durch praktische Herausforderungen erschwert. In Abschnitt 3.2.2 wurde deutlich, dass sich kontrafaktische Erklärungen meist nicht auf eine ideale Intervention stützen: Zum einen sind viele moderne KI-Systeme so komplex, dass eine Erklärung nicht sinnvoll alle einflussreichen Faktoren benennen kann. Dieselbe Problematik besteht allerdings auch bei z. B. naturwissenschaftlichen Explananda – Woodwards *ideale* Intervention ist auch als Ideal zu verstehen. Problematischer ist, zum anderen, dass kontrafaktische Erklärungen Handlungen anleiten sollen. Da die Korrelationen innerhalb des KI-Systems nicht zwangsläufig die kausalen Zusammenhänge in der Welt widerspiegeln, geben kontrafaktische Erklärungen nicht immer sinnvolle Handlungsempfehlungen ab.

Pragmatische Erklärungen

Anders als die zuvor diskutierten Ansätze fassen pragmatische Theorien Kontextfaktoren als elementaren Bestandteil von Erklärungen auf. Eine einflussreiche pragmatische Theorie

der Erklärung ist neben jener von Peter Achinstein (1983) der Ansatz von Bas van Fraassen (1980). Van Fraassen beginnt mit Aristoteles' Beobachtung, dass Erklärungen Antworten auf Warum-Fragen sind. Ob eine solche Antwort als Erklärung gelten kann, hängt davon ab, ob sie für die Adressatin der Erklärung aufschlussreich ist. Um wiederum aufschlussreich zu sein, muss diese Antwort der kontextuell vordergründigen Relevanz-Relation und Kontrastklasse der Warum-Frage Rechnung tragen. Die Relevanz-Relation bestimmt, welche Art von Information angefordert wird: In manchen Kontexten ist eine Fragestellerin vielleicht daran interessiert, die Ursachen für ein Ereignis zu erfahren; in anderen Fällen geht es ihr womöglich eher darum, zu verstehen, warum das Ereignis notwendigerweise eintreten musste. Die Kontrastklasse beschreibt den meist impliziten Kontrast, den eine Warum-Frage beinhaltet. Die Frage ‚Warum springt Willi in den Pool?‘ kann implizit z. B. danach fragen, warum *Willi* (anstatt Enno) in den Pool springt, warum er in den Pool *springt* (anstatt die Leiter zu nutzen) oder warum er *in den Pool* (anstatt in den See) springt. Welche Informationen in eine pragmatische Erklärung integriert werden, hängt also nicht allein von den (gesetzmäßigen, kausalen, mechanischen...) Gegebenheiten in der Welt ab, sondern auch davon, was für die Fragestellerin in einem gegebenen Kontext von Bedeutung ist.

Spätestens seit einem einflussreichen Aufsatz von Tim Miller (2019) ist der Grundgedanke pragmatischer Erklärungstheorien auch in XAI populär. Miller versucht darin, einen Abriss über philosophische, kognitionswissenschaftliche und psychologische Erkenntnisse zum Thema Erklärung zu bieten. Als Kerneinsicht hebt er die Kontextualität von Erklärungen hervor (ebd., S. 3) und diskutiert eine Reihe verschiedener Aspekte, die dazu beitragen, dass eine Erklärung in Bezug auf einen bestimmten Kontext eine *gute* Erklärung ist. Insbesondere die Kontrastivität von Erklärungen (siehe dazu auch P. Lipton, 1990) ist bei Miller, aber auch in der sonstigen XAI-Community, auf fruchtbaren Boden gefallen (etwa Mittelstadt et al., 2019; Nauta et al., 2023; Stepin et al., 2021). Beispielsweise gelten kontrafaktische Erklärungen als kontrastiv, wenngleich beides streng genommen nicht gleichgesetzt werden kann.³⁰

Angesichts der Vielfalt der XAI-Kontexte und der möglichen Adressat:innen von XAI-Erklärungen ist es zweifelsohne wichtig, dabei auch pragmatische Faktoren zu berücksichtigen. Im Vergleich zu vielen anderen Theorien der Erklärung bleibt van Fraassens Ansatz aber eher unkonkret. Denn ‚pragmatisch‘ heißt die Theorie daher, weil van Fraassen annimmt, dass die Bedeutung einer Warum-Frage erst durch kontextuelle Faktoren festgelegt wird. So wie sich das Pronomen ‚Ich‘ je nach Sprecher:in auf unterschiedliche

³⁰ Andrés Páez arbeitet den Unterschied heraus: Kontrastive Erklärungen sind faktiv. Eine explanatorische Antwort auf die Frage ‚Warum sprang Willi in den Pool anstatt in den See?‘ – vielleicht, ‚weil das Seewasser zu kalt war‘ – ist also wahr oder korrekt. Das hingegen, was innerhalb von XAI häufig als kontrafaktische Erklärung bezeichnet wird, ist streng genommen gar keine Erklärung. Vielmehr handelt es sich um eine hypothetische Aussage, die einen anderen Verlauf der Geschehnisse annimmt: ‚Was hätte anders sein müssen, damit Willi in den Pool anstatt den See gesprungen wäre?‘ – ‚Das Wasser im See hätte mindestens so warm sein müssen wie das im Pool.‘ (Páez, 2019, Fn. 17 auf S. 454).

Personen bezieht, wird die Relevanz-Relation und Kontrastklasse einer Warum-Frage erst durch den Kontext spezifiziert, in dem sie geäußert wird. Entsprechend existieren keine allgemeingültige Form oder kontextübergreifende Inhalte von Erklärungen (van Fraassen, 1980, S. 155–156). Die pragmatische Theorie der Erklärung gibt XAI-Forscher:innen daher wenig konkrete Anhaltspunkte dafür an die Hand, wie eine XAI-Erklärung aussehen könnte.³¹

Gründe-Erklärungen

Die letzte Theorie der Erklärung, die ich betrachten möchte, stammt nicht aus der Wissenschafts- sondern aus der Handlungstheorie: Gründe-Erklärungen (Alvarez, 2010; Hieronymi, 2011). Während sich die zuvor diskutierten Theorien vornehmlich auf wissenschaftliche Kontexte beziehen, adressieren Gründe-Erklärungen alltägliche menschliche Handlungen. Zum Beispiel: ‚Warum machte Willi einen Salto in den Pool?‘ ‚Weil sein Schwarm Enno zuschaute.‘ Häufig werden drei Arten (oder Redeweisen) von Gründen unterschieden: Normative, motivierende und explanatorische Gründe (Alvarez & Way, 2024). *Normative Gründe* sind Fakten, die objektiv für oder gegen eine Handlung sprechen. Zusammengekommen machen normative Gründe eine Handlung richtig oder falsch. Übertragen auf unser Beispiel: Dass sein Schwarm Enno zuschaut, spricht dafür, dass Willi einen Salto in den Pool macht. *Motivierende Gründe* sind Erwägungen, im Lichte derer eine Person handelt. Zum Beispiel: Dass sein Schwarm Enno zuschaut, motiviert Willi dazu, einen Salto in den Pool zu machen. *Explanatorische Gründe* erklären eine Handlung: Dass sein Schwarm Enno zuschaut, erklärt, warum Willi einen Salto in den Pool macht. Normative Gründe stimmen nicht immer mit motivierenden bzw. explanatorischen Gründen überein. Dass Enno (scheinbar) zuschaut, motiviert Willi vielleicht zu seinem Salto. Es kann aber sein, dass Enno in Wirklichkeit in sein Buch vertieft ist und sein (vorgebliches) Zuschauen daher keinen normativen Grund dafür darstellt, einen Salto in den Pool zu machen.

Auch Gründe-Erklärungen wurden auf den KI-Kontext übertragen (Baum et al., 2022; de Graaf & Malle, 2017; Schmidt, 2022). Beispielsweise schlagen Kevin Baum et al. vor, dass Gründe-Erklärungen besonders gut geeignet sind, um die moralische Verantwortlichkeit von Menschen in der Schleife bei KI-gestützten Entscheidungen sicherzustellen (Baum et al., 2022). Die einflussreichen Aspekte der Eingabedaten entsprechen den explanatorischen Gründen des Systems für eine bestimmte Ausgabe. Diese können dann in Form einer Gründe-Erklärung angegeben werden (ebd., S. 18). Besonders nützlich ist dabei die Unterscheidung zwischen normativen und motivierenden bzw. explanatorischen Gründen. Denn sie hebt hervor, dass die motivierenden Gründe eines Systems für einen Output nicht zwangsläufig objektiv für diese Ausgabe sprechen – auch KI-Systeme machen bekanntlich

³¹Zudem ist die für van Fraassens Ansatz so zentrale Kontrastivität von Erklärungen philosophisch umstritten. Kritiker:innen bezweifeln, dass erklärungsverlangende Warum-Fragen immer eine (implizite) Kontrastklasse aufweisen. Mitunter lassen sich solche Fragen auch problemlos verstehen (und sinnvoll beantworten), ohne zuvor die Kontrastklasse zu bestimmen.

Fehler.

Da Gründe-Erklärungen uns allen aus alltäglichen Kontexten vertraut sind, könnten sie insbesondere für Nichtfachleute intuitiver nachvollziehbar sein als naturwissenschaftliche Erklärungen. Allerdings ist offensichtlich, dass KI-Systeme nicht wirklich aus Gründen handeln und nur im übertragenen Sinne durch Gründe-Erklärungen erklärbar sind. Baum et al. nehmen daher ein funktionalistisches Bild von Gründen an, das nicht voraussetzt, dass das KI-System ein Bewusstsein besitzt (ebd.). Auch hier zeigt sich also eine Diskrepanz zwischen existierenden Erklärungstheorien und den spezifischen Anforderungen, die sich in XAI-Kontexten ergeben.

Erklärbar oder verständlich?

In der Gesamtschau treten bei der Anwendung vieler philosophischer Erklärungstheorien auf KI-Systeme zwei wiederkehrende Schwierigkeiten auf: Erstens wurden gängige Theorien der Erklärung für gänzlich andere Explananda entwickelt (zumeist natürliche Phänomene, aber auch menschliche Handlungen). Das führt dazu, dass sie nur mit signifikanten Anpassungen auf KI-Systeme übertragen werden können. Die resultierenden Erklärungen beruhen z. B. häufig auf Metaphern bzw. Analogien, welche die Faktizität der Erklärung gefährden (vgl. Páez, 2026, S. 143). Zweitens: Selbst wenn genuine (z. B. mathematische) Erklärungen für KI-Systeme generiert werden könnten, wäre bei den meisten nicht-pragmatischen Theorien dadurch noch nicht sichergestellt, dass die erklärten Aspekte auch verständlich gemacht werden. Damit XAI sein Ziel erreicht, braucht es jedoch nicht bloß eine akkurate Darstellung explanatorischer Zusammenhänge, sondern eine Repräsentation, welche die epistemische Opazität von KI für relevante Stakeholder hinreichend reduziert.

Angesichts dieser beiden Herausforderungen möchte ich mich nachfolgend auch auf den Begriff des Verstehensermöglichers stützen. In Abschnitt 3.1 habe ich Verstehensermöglicher als Repräsentationen systembezogener Aspekte charakterisiert, die explanatorisch sein können, aber nicht müssen. Dieser Begriff vermeidet beide Schwierigkeiten: Zum einen lässt er zu, dass KI-Systeme oder einzelne ihrer Aspekte verständlich gemacht werden können, auch wenn sie im philosophischen Sinne nicht erklärbar sein sollten. Verstehensermöglicher schließen neben Erklärungen auch andere Repräsentationen ein, die Aufschluss über relevante systembezogene Aspekte geben. Dadurch muss ich mich nicht darauf festlegen, dass Erklärungen das einzige oder beste Mittel sind, um KI-Systeme bzw. einzelne ihrer Aspekte verständlich zu machen. Zum anderen betont der Begriff des Verstehensermöglichers das epistemische Ziel, das in XAI-Kontexten erreicht werden soll. Dadurch tritt die Frage in den Hintergrund, ob KI-Systeme erklärbar sind bzw. ob XAI-Erklärungen den Kriterien gerecht werden, die philosophische Theorien der Erklärung formulieren. Stattdessen rückt die Frage danach in den Vordergrund, ob die Repräsentationen, mit denen in XAI-Kontexten Verstehen hervorgebracht werden soll, diesem Ziel zuträglich sind (vgl. Malfatti, 2023, Kap. 4; Páez, 2019, 2026).

3.3.2 Sind Erklärungen notwendig für das Verstehen von KI-Systemen?

Es ist offensichtlich, dass Erklärungen ein wichtiges und gängiges Mittel dafür sind, um Verstehen zu erwerben. Dinge, die wir trotz großer Bemühungen nicht verstehen, bezeichnen wir als ‚unerklärlich‘. Wenn jemandem etwas unverständlich ist, fehlt dieser Person typischerweise eine Erklärung. Und vermutlich wird sich eine solche Person darum bemühen, eine Erklärung zu finden, um ihr Unverständnis zu beseitigen (Malfatti, 2023, S. 95). Vor diesem Hintergrund hat sich eine Subdebatte speziell dazu entfaltet, ob Erklärungen womöglich sogar notwendig für Verstehen sind (siehe ebd., Kap. 4.4, für eine Übersicht). Einige Autor:innen verteidigen die Notwendigkeit von Erklärungen für den Verstehenserwerb (Khalifa, 2017; Strevens, 2013; vgl. auch Grimm, 2014), oder nehmen zumindest an, dass ideales oder maximales Verstehen Wissen von Erklärungen erfordert (Kelp, 2015). Eine Reihe anderer Autor:innen hat hingegen argumentiert, dass Verstehen auch auf anderem Wege erworben werden kann (Baumberger, 2019; Dellsén, 2020; Elgin, 2017, 2021; Gijsbers, 2013; Kelp, 2021; Kvanvig, 2009b; P. Lipton, 2009; Malfatti, 2023). In diesem Abschnitt vertrete ich die Position, dass die letztgenannten Autor:innen Recht haben und dass Erklärungen auch in KI-Kontexten nicht notwendig für Verstehen sind.

Warum ist das relevant? Erinnern wir uns noch einmal an Abschnitt 3.3.1. Dort habe ich zweierlei aufgezeigt: Zum einen ist unklar, ob KI-Systeme im philosophischen Sinne erklärbar sind. Zum anderen ist das Vorhandensein einer Erklärung den meisten Erklärungstheorien zufolge nicht hinreichend für den Erwerb von Verstehen. Mein Vorschlag lautete daher, in XAI-Kontexten (auch) Verstehensmöglichkeiten in den Blick zu nehmen – also Repräsentationen systembezogener Aspekte, die nicht notwendigerweise explanatorisch sind. Dieser Vorschlag setzt natürlich voraus, dass Verstehen auf anderem Wege als durch Erklärungen erworben werden kann.

In der Literatur werden viele verschiedene Fälle von Verstehen ohne Erklärung diskutiert, von denen einige auch für XAI-Kontexte relevant sind. Nachfolgend übertrage ich diese Beispiele auf das Verstehen von KI-Systemen. Wenn meine Argumentation erfolgreich ist, sind Erklärungen nicht notwendig, um KI-Systeme oder einzelne ihrer Aspekte verständlich zu machen. Das stützt wiederum meine Position, neben Erklärungen auch andere Verstehensmöglichkeiten in den Blick zu nehmen.

Nicht-propositionale Repräsentationen

Beginnen wir mit einem Fall aus einem vieldiskutierten Aufsatz von Peter Lipton, in dem er mehrere Beispiele für den Erwerb von Verstehen ohne Erklärungen entwickelt. Eine Gruppe von Beispielen betrifft nicht-propositionale Mittel des Verstehenserwerbs. Unter anderem schlägt Lipton vor, dass man durch ein Orrery (also ein mechanisches Modell unseres Sonnensystems) oder eine Visualisierung verstehen könne, warum manche Planeten von der Erde aus betrachtet scheinbar die Bewegungsrichtung auf ihrer Umlaufbahn wechseln, ohne dieses Phänomen erklären zu können (P. Lipton, 2009, S. 45).

Die Diskrepanz zwischen Verstehen und Erklärung ergibt sich Lipton zufolge daraus, dass Erklärungen stets propositional und explizit seien (ebd., S. 43). Im Gegensatz dazu könne man mithilfe eines Orrerys die scheinbar retrograde Bewegung eines Planeten implizit verstehen, ohne das eigene Verständnis in Form einer expliziten, propositionalen Erklärung verbalisieren zu können (vgl. ebd., S. 45–46). Daher, so Liptons Ergebnis, seien Erklärungen nicht notwendig für Verstehen.

Wie in Abschnitt 3.2 deutlich wurde, kommen auch in XAI-Kontexten häufig nicht-propositionale Repräsentationen zum Einsatz. Neben Salienzkarten (siehe etwa Abbildung 3.2b) spielt beispielsweise auch die Datenvisualisierung in Form von Diagrammen oder Graphen eine große Rolle für das Verstehen bestimmter Aspekte von KI-Systemen (vgl. Skiena, 2017, Kap. 6). Wir können uns Fälle vorstellen, in denen jemand mithilfe einer solchen Repräsentation Aspekte von KI-Systemen versteht, ohne aber das eigene Verständnis artikulieren zu können. Handelt es sich dabei also um Fälle von Verstehen ohne Erklärung?

Bei genauerer Betrachtung lässt sich an verschiedenen Stellen von Liptons Argumentation einhaken. Beispielsweise ist unklar, welche Rolle der suggerierte Zusammenhang zwischen Propositionalität und Explizitheit für Liptons Argumentation spielt. Lipton scheint anzunehmen, dass nicht-propositionale Repräsentationen (und damit Nicht-Erklärungen) eher dazu angetan sind, implizites Verstehen hervorzubringen. Allerdings ist die Verwendung einer nicht-propositionalen Repräsentation *prima facie* weder notwendig noch hinreichend für das Ausbilden impliziter Überzeugungen. Ich kann z. B. auf Grundlage eines nicht-propositionalen Orrerys explizit glauben, dass der Jupiter der größte Planet unseres Sonnensystems ist. Umgekehrt kann man anhand einer propositionalen Beschreibung zu der impliziten Überzeugung gelangen, dass es sich beim Merkur um den sonnennächsten Planeten handelt. (Ich komme in Abschnitt 7.3 auf den Übergang von nicht-propositionalen Repräsentationen zu propositionalen Überzeugungen zurück.) Daher scheint es sinnvoll, die Frage nach der Implizitheit von Verstehen getrennt von jener nach der Propositionalität von Erklärungen zu betrachten.

Aus dieser Perspektive könnte man, erstens, im Anschluss an Linda Zagzebski argumentieren, dass Verstehen nicht implizit sein kann (Zagzebski, 2001, S. 246).³² Wenn man die retrograde Bewegung der Planeten nach Betrachtung oder Manipulation des Orrerys nur implizit erfasst, würde man demnach kein Verstehen besitzen. Zweitens ließe sich anzweifeln, dass Erklärungen notwendigerweise propositional sind. Demnach könnte auch das Orrery als genuine Erklärung gelten.³³ Ich finde eine dritte Option am vielversprechends-

³²Zagzebski nimmt an, dass Verstehen notwendigerweise transparent ist: „Understanding [...] is a state that is constituted by a type of conscious transparency. It may be possible to know without knowing that one knows, but it is impossible to understand without understanding that one understands“ (Zagzebski, 2001, S. 246). Sie vertritt also eine Art ‚UU-Prinzip‘ (analog zum KK-Prinzip für Wissen; ‚KK‘ steht für ‚*knowing that one knows*‘). Ich denke, dass Verstehen nicht notwendigerweise vollkommen transparent ist (wie auch Gordon, o. D., Kap. 5.a.; Grimm, 2006, S. 517–518; Grimm, 2017; Hills, 2016, S. 663; Hubert & Malfatti, 2023, Kap. 5.3; Pritchard, 2010, S. 75–76), wenngleich Verstehen meinem Vorschlag zufolge ein internalistisches Element aufweist (mehr dazu in Abschnitt 10.2.2).

³³Manche Philosoph:innen merken (beiläufig) an, dass nicht-propositionale Repräsentationen als Erklärungen

ten: Man könnte Liptons Behauptung zurückweisen, dass Erklärungen notwendigerweise explizit sind (Gijsbers, 2013, S. 518; Strevens, 2013, S. 512). Nur weil jemand die explanatorischen Zusammenhänge, welche die scheinbar retrograde Bewegung der Planeten erklären, nicht artikulieren kann, heißt das nicht, dass diese Person keine Erklärung besitzt. Dieser Argumentationslinie zufolge fungiert nicht das Orrery als Erklärung, sondern es dient als Mittel, um eine implizite Erklärung (mental) zu konstruieren. Auch das Verstehen im Orrery-Fall würde demnach durch eine Erklärung zustande kommen. Kurz: Nur weil jemand einen Verstehensermöglicher verwendet, der keine Erklärung ist, beweist dies noch nicht, dass es sich um einen Fall von Verstehen ohne Erklärung handelt.

Räumliche Zusammenhänge

Dennoch sind nicht-propositionale Repräsentationen ein guter Ansatzpunkt, um Fälle von Verstehen ohne Erklärung zu finden. Denn solche Repräsentationen werden häufig dafür genutzt, um nicht-explanatorische Zusammenhänge verständlich zu machen. Ein Beispiel aus der Wissenschaftsgeschichte ist die Entdeckung der Doppelhelix-Struktur der DNA durch James Watson and Francis Crick mithilfe eines Modells aus Pappe.³⁴ Die Wissenschaftler verwendeten dieses Modell, um physisch auszuprobieren, wie die DNA aufgebaut sein könnte. Der Überlieferung zufolge kam Watson beim Zurechtschieben der Papp-Moleküle die entscheidende Idee der Doppelhelix (vgl. dazu NCBI, o. D.).

Inwiefern handelt es sich hierbei um einen Fall von Verstehen ohne Erklärung? Watson und Crick waren sicherlich dazu in der Lage, (bis zu einem gewissen Grad) zu erklären, warum die DNA die Form einer Doppelhelix aufweist, und konnten mithilfe der Struktur Erklärungen für verschiedene Phänomene formulieren. Doch betrachten wir stattdessen jemanden ohne Expertise in der Molekularbiologie. Eine solche Person kann mithilfe eines physischen Modells (oder einer Visualisierung) plausiblerweise die Struktur des Moleküls verstehen – also die räumliche Anordnung der vier verschiedenen Nukleinbasen und der in sich verdrehten Rückgrat-Stränge. Aber diese Person kann vermutlich nicht erklären, warum die DNA diese Doppelhelix-Struktur aufweist, oder auf dieser Grundlage weiterführende mikrobiologische Erklärungen formulieren. Räumliche Zusammenhänge scheinen also auch ohne Erklärung verstanden werden zu können – ganz einfach, weil diese Zusammenhänge nicht zwangsläufig explanatorisch sind.

Ein solches Verständnis kann auch in XAI-Kontexten relevant sein. Beispielsweise kann sich die räumliche Anordnung von Features auf die Klassifikation eines Bildes auswirken.

in Frage kommen, bleiben aber eine Theorie der nicht-propositionalen Erklärung schuldig (etwa Craver, 2014, S. 32; de Regt, 2015, Fn. 18 auf S. 3793; Erasmus et al., 2021, S. 851; Riggs, 2003, S. 218 oder Ylikoski, 2009, S. 114). Einzelne Autor:innen üben ausdrücklich Kritik daran, dass Philosoph:innen visuellen Erklärungen oder generell visuellen, epistemisch relevanten Repräsentationen zu wenig Aufmerksamkeit schenken (z. B. Bechtel & Abrahamsen, 2005; Kozak, 2023; Perini, 2005a). Der Standardauffassung zufolge sind Erklärungen allerdings propositional. Vgl. etwa J. Kim, 1994, S. 54; Kozak, 2023, S. 150; P. Lipton, 2009, S. 43; Malfatti, 2023, S. 125; McCain, 2016, S. 137; Strevens, 2013, S. 510 oder Woodward, 2003, S. 202.

³⁴Vielen Dank an Lena Kästner, die mich auf dieses schöne Beispiel hingewiesen hat.

Johannes Rabold et al. illustrieren dies am Beispiel der Bildklassen ‚Haus‘ und ‚Turm‘: Für beide Klassen könnte die Detektion von Fenstern eine Rolle spielen. Es ist aber denkbar, dass im Falle der Turm-Klasse entscheidend ist, dass die Fenster übereinander sind, während Häuser in der Regel an mehreren nebeneinander liegenden Fenstern erkannt werden (Rabold et al., 2020a, S. 182). Um zu verstehen, warum ein Bildklassifikationssystem ein bestimmtes Foto als ‚Turm‘ klassifiziert hat, kann daher ein Verständnis räumlicher Zusammenhänge erforderlich sein. Dieses Verständnis muss wiederum nicht notwendigerweise durch Erklärungen hervorgebracht werden.

Ein anderes Beispiel ist der räumliche Aufbau von neuronalen Netzen, der sich ebenfalls besonders gut mithilfe von Visualisierungen verstehen lässt (siehe Abbildung 2.3a). Noch besser fügt sich dieser Fall aber in die folgende Kategorie von Beispielen für Verstehen ohne Erklärungen ein.

Metaphern

In Abschnitt 2.1.2 habe ich insistiert, dass künstliche neuronale Netze keine genuinen Netzwerke sind und dass darin streng genommen kein ‚Signal‘ durch schichtweise angeordnete ‚Neuronen‘ und ‚Kanten‘ wandert. Vielmehr ist das Netzwerk eine nützliche Metapher, um die Funktionsweise dieser Modelle verständlich zu machen. Gleichmaßen habe ich in Abschnitt 3.3.1 viele Versuche, bestehende Theorien der Erklärung auf KI-Systeme anzuwenden, aufgrund des metaphorischen Charakters der resultierenden ‚Erklärungen‘ kritisiert. Wenn man von den ‚Ursachen‘ für eine Klassifikation spricht, oder die ‚motivierenden Gründe‘ eines Systems anführt, obwohl man sich auf rein mathematische Zusammenhänge bezieht, ist fraglich, ob es sich dabei um eine genuine Erklärung im Sinne der zugrundeliegenden Erklärungstheorie handelt.

Gleichzeitig ist es plausibel, dass solche Analogien bzw. Metaphern Verstehen hervorbringen können, auch ohne explanatorisch zu sein (vgl. P. Lipton, 2009, S. 53; Malfatti, 2023, Kap. 4.4.3; Páez, 2026, S. 143). Das Merriam-Webster-Wörterbuch definiert *Metapher* wie folgt: „a figure of speech in which a word or phrase literally denoting one kind of object or idea is used in place of another to suggest a likeness or analogy between them (as in *swimming in paperwork*)“ (Hervorhebung im Original). Metaphern können also eine Ähnlichkeit zwischen Dingen oder Ideen herausarbeiten. Im Falle der oben beschriebenen Ursachen- oder Gründe-Analogie ist die Ähnlichkeit vermutlich funktional (vgl. Baum et al., 2022, S. 18), die Netzwerk-Metapher weist vornehmlich auf eine strukturelle Ähnlichkeit hin. Der entscheidende Punkt ist, dass solche Metaphern nicht (zwangsläufig) explanatorisch sind und damit mögliche Beispiele für Verstehen ohne Erklärung darstellen.

Unifikation

Betrachten wir als nächstes einen Fall von Verstehen ohne Erklärung, der *prima facie* dem Verstehen von KI-Systemen ähnelt, bei genauerer Betrachtung aber relevante Un-

terschiede aufweist. Victor Gijsbers führt ein biologisches Klassifikationssystem aus dem 18. Jahrhundert als Beispiel für Verstehen ohne Erklärung an (Gijsbers, 2013). Dieses Klassifikationsschema weist jedem Organismus auf Basis einer möglichst geringen Anzahl an Merkmalen (z. B. die Schädelform) einen einzigen Platz zu (z. B. die Kategorie der Warmblüter). Das Schema erklärt nicht, warum Warmblüter eine bestimmte Schädelform haben, aber es ordnet die biologische Welt und zeigt systematische Gemeinsamkeiten und Unterschiede zwischen den Spezies auf. Gijsbers unterscheidet daher Verstehen durch Erklärungen von Verstehen durch Unifikation.

Doch warum sollte es sich bei Verstehen durch Unifikation um einen Fall von Verstehen ohne Erklärung handeln? Schließlich existieren mehrere unifikationistische Erklärungstheorien (unter anderem von Philip Kitcher, Gerhard Schurz und Karel Lambert). Deren Grundidee lautet, dass gute wissenschaftliche Erklärungen und Theorien eine möglichst große Anzahl verschiedener Phänomene unter sich vereinigen, indem sie zeigen, dass all diese Phänomene durch eine möglichst kleine Anzahl derselben Prinzipien hervorgebracht werden.³⁵ Beispielsweise lassen sich mithilfe der Newton'schen Mechanik nicht nur die Bewegungen von diversen Objekten auf der Erde erklären, sondern auch von Planeten und anderen Himmelskörpern. Newtons Mechanik ist damit vereinigender als die aristotelische Annahme, dass auf der Erde und im All unterschiedliche Gesetze gelten (ebd., S. 517).

Gijsbers grenzt hingegen Erklärungen und Unifikation voneinander ab (ebd., Kap. 2). Unifikation sei nicht notwendig für Erklärungen, da eine Erklärung nur so vereinigend sein könne, wie es die Wirklichkeit zulasse. Wenn z. B. im All andere Gesetze gelten würden als auf der Erde, wäre die aristotelische Position der Newton'schen Theorie überlegen – obwohl Newtons Mechanik vereinigender ist als Aristoteles' Bewegungslehre. Unifikation sei auch nicht hinreichend für Erklärungen. Gijsbers' Beispiel sind die (durchaus unifikatorischen) Gesetze der Optik, mit deren Hilfe sich die Länge eines Schattens deduzieren, aber nicht erklären lässt. Ein anderes Beispiel sind übergeordnete mathematische Gleichungen, die in sehr verschiedenen Teilgebieten der Physik zum Einsatz kommen. Nur weil dieselben Gleichungen angewandt werden können, ergibt sich daraus noch keine gemeinsame, vereinigende Erklärung z. B. eines mechanischen und elektrischen Systems (Woodward, 2014, S. 208).

Nehmen wir an, dass Gijsbers Recht hat und es sich bei dem oben beschriebenen biologischen Klassifikationsschema um einen Fall nicht-explanatorischer Unifikation handelt. Angesichts dieses Beispiels könnte man denken, dass ML-basierte Systeme sogar *ausschließlich* durch Unifikation verstanden werden können: Stellen wir uns vor, dass ein Bildklassifikationssystem Katzenbilder aufgrund der spitzen Form der Ohren als ‚Katze‘ klassifiziert. In diesem Fall haben sich spitze Ohren während des Trainings als Merkmal

³⁵Aus Platzgründen konnte ich in Abschnitt 3.3.1 nicht näher auf diese (und viele andere) Theorien der Erklärung eingehen. Es soll aber nicht unerwähnt bleiben, dass auch unifikationistische Erklärungen auf KI-Systeme übertragen wurden. Beispielsweise argumentiert Yunus Prasetya, dass neuronale Netze (unifikationistisch) nicht erklärbar sind, weil diese Erklärungen ein möglichst einfaches Explanans voraussetzen (Prasetya, 2022).

erwiesen, das mit der Bildklasse ‚Katze‘ korreliert und sich daher für die Zwecke der Klassifikation eignet. Diese Korrelation erklärt aber nicht, warum etwas ein Katzenbild ist. Noch deutlicher sind die Parallelen bei unüberwachten Lernverfahren wie der Clusteranalyse (siehe Abschnitt 2.2.2). Die Clusteranalyse teilt Datenpunkte anhand von Gemeinsamkeiten in Gruppen ein – es handelt sich um ein ähnlichkeitsbasiertes Klassifikationsschema, das die Ähnlichkeiten nicht erklärt (Knüsel & Baumberger, 2020, S. 55).

Allerdings gibt es einen entscheidenden Unterschied zu Gijsbers’ Beispiel: XAI versucht typischerweise, *KI-Systeme* verständlich zu machen. Und die Detektion einer bestimmten Ohrenform erklärt durchaus, warum das System ein Bild als ‚Katze‘ klassifiziert, oder warum es bestimmte Datenpunkte zu einem Cluster zusammenfasst. Man könnte das System aber prinzipiell auch verwenden, um etwas über Katzen oder Katzenbilder herauszufinden. Zwar würden spitze Ohren dann nicht erklären, warum etwas eine Katze bzw. ein Katzenbild ist. Doch wenn Gijsbers Recht hat, würde die Korrelation zwischen Ohrenform und Kategorie trotzdem einen gewissen Verstehensgrad von Katzen(bildern) herbeiführen. Hier kommt es also darauf an, welcher Repräsentationsebene das Objekt des Verstehens angehört: Geht es um das Verstehen des KI-Systems, oder um das Verstehen eines Aspekts der Welt? Ich komme in Kapitel 6 auf diese wichtige Unterscheidung zurück.

Explanatorische Irrelevanz

Ein letztes Beispiel für Verstehen ohne Erklärung, das auch für XAI-Kontexte relevant ist, stammt von Finnur Dellsén. Dellsén adaptiert diesen Fall wiederum von Peter Liptons Diskussion von Galileo Galileis berühmtem Gedankenexperiment zum freien Fall. Mit diesem Gedankenexperiment widerlegte Galilei die aristotelische Annahme, dass schwere Körper schneller fallen als leichte: Wenn man einen großen und kleinen Stein zusammenbindet und das Konglomerat von einem Turm wirft, dann müsste der kleinere Stein Aristoteles zufolge den größeren ‚ausbremsen‘ – gleichzeitig ist das Konglomerat insgesamt schwerer als der große Stein allein und müsste deswegen schneller fallen. Aristoteles’ Annahme führt also zu einem Widerspruch. Lipton zufolge bringt diese *Reductio ad absurdum* Verstehen hervor, indem sie Notwendigkeit aufzeigt, ohne aber eine Erklärung zu sein (P. Lipton, 2009, S. 46–49). Dellsén schlägt eine andere Interpretation dieses Beispiels vor: Galileos Gedankenexperiment erlaube es zu begreifen, dass zwei Faktoren (Masse und Fallgeschwindigkeit) unabhängig voneinander sind (Dellsén, 2020, Kap. 7). In diesem Fall erwirbt man Verstehen, indem man erfährt, dass etwas *nicht* explanatorisch ist (ebd., S. 1278).

Auch das Verstehen von KI-Systemen kann durch das Aufzeigen von Unabhängigkeit bzw. explanatorischer Irrelevanz verbessert werden. Beispielsweise versteht man die Ausgabe eines KI-Systems plausiblerweise besser, wenn man erfährt, welche Faktoren *keinen* Einfluss auf die Ausgabe hatten. In der Tat kann es sich dabei um eine entscheidende Information handeln: Stellen wir uns eine Richterin vor, die COMPAS benutzt, um über die erneute Inhaftierung einer Straftäterin zu entscheiden. COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) ist ein System, das die Rückfallwahrscheinlichkeit von

Straftäter:innen berechnen soll und in mehreren Bundesstaaten der USA vor Gericht, beispielsweise im Rahmen der Strafzumessung, verwendet wird. Wenn die Richterin erfährt, dass die ethnische Zugehörigkeit der Straftäterin keinen Einfluss auf die von COMPAS prognostizierte Rückfallwahrscheinlichkeit hatte, hilft ihr dies plausiblerweise dabei, die Vorhersage von COMPAS zu verstehen und beispielsweise zu beurteilen, ob das System fair ist.³⁶

Nicht alle Verstehensermöglicher sind Erklärungen

Im vorherigen Abschnitt 3.3.1 wurde zweierlei deutlich: Erstens ist unklar, ob genuine Erklärungen für KI-Systeme erlangt werden können. Zweitens sind Erklärungen in der Regel nicht hinreichend dafür, um das epistemische Ziel von XAI – also das Hervorbringen von Verstehen – zu erreichen. In diesem Abschnitt habe ich darüber hinaus argumentiert, dass Erklärungen auch nicht notwendig für den Verstehenserwerb in XAI-Kontexten sind.

Die überzeugenden Beispiele für Verstehen ohne Erklärung zeichnen sich dadurch aus, dass sie das Verstehen von Zusammenhängen ermöglichen, die nicht explanatorisch sind: seien es räumliche Zusammenhänge wie im Fall des DNA-Pappmodells, Ähnlichkeiten wie bei der Netzwerk-Metapher und der Unifikation unter gemeinsame Prinzipien oder auch der ‚Nicht-Zusammenhang‘ der explanatorischen Irrelevanz. Viele weitere Zusammenhänge sind denkbar; etwa zeitliche, mereologische, supervenierende, modale usw. Das Begreifen von Zusammenhängen ist wiederum ein charakteristisches Merkmal von Verstehen (ich komme in Abschnitt 6.1.3 auf diesen Aspekt zurück). Zusammengenommen untermauern diese Überlegungen meinen Vorschlag, im Kontext von XAI nicht bloß von Erklärungen, sondern allgemeiner von Verstehensermöglichern zu sprechen.

Verstehensermöglicher sind z. B. linguistische oder visuelle Repräsentationen, die einem Stakeholder epistemischen Zugang zu systembezogenen Aspekten verschaffen. Zwar lässt der Begriff zu, dass Verstehensermöglicher explanatorische Informationen zum Inhalt haben. Doch darüber hinaus kommen auch nicht-explanatorische Informationen in Frage, beispielsweise Informationen über räumliche Relationen, Ähnlichkeiten oder explanatorische Irrelevanz. Außerdem kommen auch nicht-propositionale Repräsentationen (z. B. Salienzkarten) als Verstehensermöglicher in Frage (vgl. auch Elgin, 2017, S. 195–196), die der Standardauffassung zufolge nicht als Erklärung gelten – wenngleich oben deutlich

³⁶Tatsächlich ist umstritten, ob COMPAS Angehörige bestimmter ethnischer Gruppen benachteiligt. Analysen von Investigativjournalist:innen der Recherche-Plattform *ProPublica* zeigen, dass die Falsch-Positiv-Rate – d. h. die Zuordnung einer mittleren oder hohen Rückfallwahrscheinlichkeit zu Personen, die tatsächlich nicht rückfällig werden – bei Schwarzen Angeklagten fast doppelt so hoch ist wie bei weißen Angeklagten (Angwin et al., 2016, o. S.). Die Autor:innen folgern daraus, dass COMPAS unfair sei, da es Schwarze Menschen systematisch benachteilige. Untersuchungen des Herstellers von COMPAS widersprechen dieser Aussage. COMPAS sei fair, da der Anteil von Personen, denen eine hohe Rückfallwahrscheinlichkeit vorausgesagt wird, bei Schwarzen und weißen Angeklagten ähnlich ist (Dieterich et al., 2016). Hier tritt das Problem zutage, dass die Fairness von KI-Systemen verschiedentlich operationalisiert werden kann, wobei diese Operationalisierungen meist nicht kompatibel miteinander sind (Kleinberg et al., 2017).

wurde, dass auch solche Repräsentationen den Ausgangspunkt für Verstehen durch (implizite, mentale) Erklärungen bieten können. Vorerst möchte ich es bei dieser allgemeinen Charakterisierung belassen, und werde auch im weiteren Verlauf dieser Arbeit keine umfassende Theorie von Verstehensermöglichere entwickeln. Allerdings werde ich in Kapitel 12 mehrere Eigenschaften von epistemisch besonders vorteilhaften Verstehensermöglichere herausarbeiten. Diese Eigenschaften sind unabhängig davon, ob ein Verstehensermöglichere als Erklärung gelten kann oder nicht.

Abschließend möchte ich hervorheben, dass Verstehensermöglichere keine exotische Kategorie darstellen. Wissenschaftler:innen innerhalb und außerhalb der KI-Forschung nutzen Visualisierungen, physische Modelle und andere nicht-propositionale Repräsentationen, um selbst Verstehen zu erwerben oder andere dabei zu unterstützen. Metaphern sind ein gängiges Mittel, um komplexe Sachverhalte zu veranschaulichen. Auch ist davon auszugehen, dass Verstehensermöglichere häufig auf ähnlichem Wege erlangt werden können wie Erklärungen: Dieselben Strategien, die für das Formulieren von Erklärungen genutzt werden (z. B. Intervention, Beobachtung oder statistische Analyse), können auch Aufschluss über explanatorische Irrelevanz geben.³⁷ Der entscheidende Punkt ist, dass Verstehensermöglichere auch nicht-explanatorische Verstehensstrategien beherbergen können und dadurch flexibler sind als der philosophische Begriff der Erklärung.

Damit sind die nötigen informatischen und philosophischen Grundlagen zu Künstlicher Intelligenz und Erklärbarkeit von KI-Systemen gelegt. Wir können nun zu Teil II dieser Arbeit übergehen, in dem ich meinen Begriff instrumentellen Verstehens entwickle.

³⁷Diesen spannenden Gedanken verdanke ich Lena Kästner.

Teil II

Instrumentelles Verstehen

4 Methodik

Im vorliegenden Teil II dieser Arbeit beantworte ich meine erste Forschungsfrage (siehe Abschnitt 1.1):

- (a) Welche Form des Verstehens ermöglicht Stakeholdern die Erfüllung gesellschaftlicher Desiderata in XAI-Kontexten?

Dieses Kapitel stellt die Methodik vor, die ich zur Beantwortung dieser Frage verwende: einen begriffsgestaltenden Ansatz (meine Übersetzung von *conceptual engineering*), der sich an Rudolf Carnaps Methode der Explikation orientiert.³⁸

Begriffsgestaltung ist ein Sammelbegriff für den Prozess der Evaluation, Verbesserung oder Neugestaltung sowie Implementierung repräsentationaler Mittel (engl. *representational devices*; gemeint sind zum Beispiel Begriffe oder Ausdrücke), um Defizite existierender repräsentationaler Mittel zu beheben und dadurch bestimmte theoretische, politische, soziale oder moralische Ziele zu erreichen (vgl. Chalmers, 2020, S. 4; Isaac et al., 2022, S. 1).³⁹ In dieser Arbeit gestalte ich einen Verstehensbegriff, welcher der von Markus Langer et al. (2021) beschriebenen Mediatorrolle von Verstehen in XAI-Kontexten gerecht wird (siehe Kapitel 1). Ich nenne diese besondere Form des Verstehens ‚instrumentelles Verstehen‘.

Das Kapitel gliedert sich wie folgt: In Abschnitt 4.1 charakterisiere und motiviere ich den Ansatz der Begriffsgestaltung im Allgemeinen und Carnaps Explikationsmethode im Besonderen. In Abschnitt 4.2 konkretisiere ich das Ziel meiner Begriffsgestaltung. Schließlich formuliere ich in Abschnitt 4.3 Adäquatheitsbedingungen für den zu gestaltenden Begriff.

4.1 Begriffsgestaltung und Explikation

Der Ansatz der Begriffsgestaltung lässt sich am besten in Abgrenzung zur klassischen Begriffsanalyse charakterisieren. Die Begriffsanalyse bemüht sich darum, möglichst präzise und akkurat die Bedeutung und Verwendung eines Begriffs zu erfassen. Üblicherweise

³⁸Die Arbeit an diesem Kapitel hat stark von dem wiederholten Austausch mit Jessica Struchhold und Marius Thienenkamp profitiert – vielen Dank an beide für hilfreiche Diskussionen und Literaturhinweise.

³⁹Meine Übersetzung erbt ein Problem des englischen Ausdrucks, da sie ausschließlich auf Begriffe verweist. Es ist nämlich umstritten, was genau eine Begriffsgestaltung gestaltet – manche Autor:innen haben argumentiert, dass dabei z. B. Sprecher:innen-Bedeutungen oder Intensionen und Extensionen von Ausdrücken gestaltet werden (Koch, 2022, Kap. 4). Vermutlich kommen eine Reihe unterschiedlicher repräsentationaler Mittel als Objekt einer Begriffsgestaltung in Frage (Isaac et al., 2022, Kap. 3). Ich lasse diese Komplikationen aus Gründen der sprachlichen Einfachheit beiseite und bleibe bei ‚Begriffsgestaltung‘.

vergleicht man bei dieser Methode die Vorhersagen, die ein Vorschlag für die Analyse eines bestimmten Begriffs ergibt, mit der intuitiven Einschätzung (hypothetischer) Beispielfälle. Sollten diese Intuitionen nicht mit den Voraussagen der fraglichen Begriffsanalyse übereinstimmen, spricht dies für eine Revision der Analyse. Der begriffsanalytische Ansatz ist also vornehmlich deskriptiv. Im Gegensatz dazu ist eine Begriffsgestaltung stärker präskriptiv. Sie zielt darauf ab, bestehende Begriffe zu verbessern oder neue Begriffe zu prägen, um damit bestimmte (z. B. theoretische) Vorteile zu erzielen.

Dieser Ansatz hat in der analytischen Philosophie in jüngerer Zeit an Popularität gewonnen – auch, weil die klassische Begriffsanalyse zunehmend kritisiert wird.⁴⁰ Die Begriffsgestaltung ist jedoch keine neue oder ungewöhnliche Methode (vgl. Greimann, 2007, Kap. 1, für einen historischen Abriss). Viele Autor:innen heben hervor, dass Philosoph:innen und Wissenschaftler:innen andauernd und immer schon repräsentationale Mittel neu- und umgestalten, wenngleich nicht immer bewusst oder mit methodischer Strenge (Chalmers, 2020; Isaac et al., 2022, S. 2–3; Koch, 2022, Kap. 3; viele Beispiele bietet Cappelen, 2018, Kap. 2).

Der übrige Abschnitt beleuchtet diese Methode und ihre Vorzüge genauer: In Abschnitt 4.1.1 motiviere ich meinen Ansatz, Forschungsfrage (a) mithilfe einer Begriffsgestaltung zu beantworten. In Abschnitt 4.1.2 begründe ich, warum ich mich spezifisch für die Carnap'sche Methode der Explikation entschieden habe. Schließlich stellt Abschnitt 4.1.3 das ‚Rezept fürs Explizieren‘ von Georg Brun (2016) vor, an dem ich mich in dieser Arbeit orientiere.

4.1.1 Warum Begriffsgestaltung?

Ich denke, dass die Methode der Begriffsgestaltung besser für die Beantwortung von Forschungsfrage (a) geeignet ist als die klassische Begriffsanalyse. Anstatt ‚die‘ Natur des Verstehens innerhalb einer allgemeingültigen Definition einzufangen, möchte ich eine bestimmte Form des Verstehens definieren, die einen spezifischen Zweck in XAI-Kontexten erfüllt – nämlich die Erfüllung gesellschaftlicher Desiderata. Dafür braucht es einen Verstehensbegriff, der dieser besonderen Rolle gerecht wird.

Diese Motivation lässt sich weiter konkretisieren. Anknüpfend an Steffen Koch können wir zwei wesentliche Argumente unterscheiden, die häufig für das Verfolgen eines begriffsgestaltenden Ansatzes angeführt werden. Beginnen wir mit dem ersten Argument. Demzufolge weisen manche oder sogar viele unserer bestehenden Begriffe problematische Defizite auf und müssen daher evaluiert, verbessert oder ersetzt werden. Da dieses Argu-

⁴⁰Beispielsweise ist die intuitive Einschätzung von Fallbeispielen, die für die Begriffsanalyse so zentral ist, viel volatiler als lange angenommen. Forschung im Bereich der experimentellen Philosophie hat z. B. ergeben, dass die relevanten Intuitionen auch von der Präsentation der Beispiele oder dem soziokulturellen Hintergrund abhängen (Koch, 2022, S. 111–112). Zudem legen die anhaltenden Debatten über so gut wie jede vorgeschlagene Begriffsanalyse nahe, dass diese Methode insgesamt betrachtet wenig erfolgreich ist (Nado, 2021, S. S1508–S1509).

ment genutzt werden kann, um grundsätzlich eine kritische Prüfung unserer Begriffe zu motivieren, nennt es Koch das ‚skeptische Argument‘ (Koch, 2022, S. 103–104; vgl. auch Cappelen, 2018, S. 6).

Koch unterscheidet drei Arten begrifflicher Defizite (Koch, 2022, Kap. 2.1; vgl. Cappelen, 2018, Kap. 2.3.1, für eine feinkörnigere Taxonomie): *Semantisch* defizitär sind beispielsweise leere, vage oder inkonsistente Begriffe (z. B. Begriffe, die zu Paradoxien führen). *Epistemisch* defizitäre Begriffe sind unzureichend für die Erreichung epistemischer Ziele, etwa, wenn sie nicht angemessen mit der Wirklichkeit korrespondieren (z. B. behaupten Befürworter:innen der Extended-Mind-Theorie, dass der Begriff *Überzeugung* auch auf Notizbücher oder Kalender ausgeweitet werden sollte, um nützlich zu sein). *Moralisch* defizitäre Begriffe fügen moralisch relevanten Personen(gruppen) Schaden zu (etwa, wenn der Begriff *Vergewaltigung* nicht-einvernehmlichen Sex zwischen Ehepartner:innen ausschließt). In der Praxis kann der Übergang zwischen diesen drei Arten von Defiziten allerdings fließend sein (Koch, 2022, S. 104).

Welche dieser Defizite könnten für mein Vorhaben relevant sein? Betrachten wir dazu einige Beispiele, wie der Verstehensbegriff in der XAI-Debatte verwendet wird. Wie in Kapitel 1 dargestellt, wird dort häufig implizit oder explizit ein Zusammenhang zwischen dem durch Erklärbarkeitsansätze generierten Verstehen und der Erfüllung ganz unterschiedlicher gesellschaftlicher Desiderata postuliert (Fleisher, 2022, S. 5; Krishnan, 2020; Langer et al., 2021):

If there is an autonomously driving car involved in an accident, it is of great interest to the developer, to the people involved and to the legal system to understand the reasons why the accident happened in order to fix the system and to sue the person responsible for the accident. (Burkart & Huber, 2021, S. 250)

[...] we risk to create and use decision systems that we do not really understand. This impacts not only information on ethics but also on accountability, on safety, and on industrial liability. (Guidotti et al., 2019, S. 93:2; Literaturangaben entfernt)

Machine-learning explanations enable users to understand how the data is processed. They aim to bring awareness to possible bias and system malfunctions. (Mohseni et al., 2020, S. 1:4)

There are now diverse use-case specific motivations for allowing humans to *understand* a given software component [...]. This starts with legal reasons [...]. Many further reasons of public interest, like fairness or security, as well as business interests like ease of debugging, knowledge retrieval, or appropriate user trust have been identified. (Schwalbe & Finzel, 2024, S. 3044; Hervorhebung im Original, Literaturangaben entfernt)

Diesen Zitaten zufolge können Erklärbarkeitsansätze unter anderem zur Fairness und Sicherheit von KI-Systemen, zu angemessenem Vertrauen und zur Wahrung der menschlichen Verantwortlichkeit beitragen, indem sie verschiedenen Personen ermöglichen, diese Systeme zu *verstehen*. Allerdings stammen die obigen Aussagen nicht von Erkenntnistheoretiker:innen, sondern zumeist von Informatiker:innen. Sie beziehen sich auf den

Alltagsbegriff von Verstehen und nutzen ihn lediglich, um die Relevanz von Erklärbarkeit zu verdeutlichen. Ihr Ziel ist (natürlich) nicht, jenes mysteriöse Verstehen näher zu fassen, das all diesen Desiderata zuträglich sein soll. Zwar gibt es mittlerweile auch diverse Arbeiten, die philosophische Überlegungen zu Verstehen mit Erklärbarkeit zusammenbringen (etwa Fleisher, 2022; Páez, 2019; Schmidt, 2026; Sullivan, 2022). Doch gibt es meines Wissens keinen philosophischen Beitrag, der diesen instrumentellen Charakter in den Mittelpunkt einer Definition von Verstehen in XAI-Kontexten rückt. In der Konsequenz haben wir es in der XAI-Literatur normalerweise mit dem nicht näher bestimmten Alltagsbegriff zu tun, wenn Verstehen als Mittel zur Erfüllung von Desiderata beschrieben wird.

Dieser Alltagsbegriff weist bis zu einem gewissen Grad alle drei Arten von Defiziten – semantisch, epistemisch und moralisch – auf, die Koch unterscheidet. Erstens ist *Verstehen*, wie die meisten alltagssprachlichen Begriffe, vage und damit semantisch defizitär. Verstehen kann unter anderem eine epistemische Errungenschaft, ein Aha-Gefühl oder Mitgefühl meinen und z. B. Menschen, Hunden oder KI-Systemen zugeschrieben werden. (Ich komme in Abschnitt 4.2.1 auf solche Bedeutungsvarianten zurück.) Wenn unklar ist, welche Form von Verstehen dazu angetan ist, die Brücke zur Erfüllung von Desiderata zu schlagen, ist das *prima facie* problematisch: Das epistemische Ziel von XAI bleibt unterbestimmt. Zweitens ist der Alltagsbegriff, zumindest im Kontext von XAI, epistemisch defizitär: Ein allgemeiner Verweis auf Verstehen lässt z. B. im Ungefähren, inwiefern das mithilfe eines Erklärbarkeitsansatzes erworbene Verständnis der Erfüllung von Desiderata zuträglich sein kann. Und drittens könnte man befürchten, dass der Alltagsbegriff moralisch defizitär ist: Offenkundig genügt es nicht, dass irgendjemand irgendetwas zu irgendeinem Grad versteht, um Desiderata zu erfüllen. Ohne einen präzisen Verstehensbegriff besteht daher das Risiko, dass moralisch relevante Desiderata unerfüllt bleiben.

Begriffliche Defizite diskutiert Koch im Kontext des skeptischen Arguments. Die wichtigste Motivation für mein Vorhaben ist aber das zweite Argument, das begriffsgestaltende Vorhaben ihm zufolge häufig motiviert: das ‚Feintuning-Argument‘. Es besagt, dass unsere Alltagsbegriffe häufig zu unpräzise für die Formulierung und Bearbeitung z. B. philosophischer oder naturwissenschaftlicher Probleme sind. Um diese Probleme effektiv lösen zu können, müssen bestehende Begriffe daher in Hinblick auf die jeweilige Zielsetzung geprüft und bei Bedarf präzisiert werden. Dem zweiten Argument zufolge sind unsere Begriffe also nicht unbedingt *grundsätzlich* defizitär, sondern vornehmlich relativ zu den Zielen der Begriffsgestalterin (Koch, 2022, S. 108–109).⁴¹

Übertragen auf den Gegenstand dieser Arbeit ist der Alltagsbegriff von Verstehen also nicht zwangsläufig unzulänglich, wenn es um Verstehen im Allgemeinen geht oder wenn wir uns in alltäglichen Kontexten auf Verstehen beziehen. Der Begriff wird allerdings problematisch, wenn man den Zusammenhang zwischen Verstehen und Desiderata in

⁴¹ Allerdings schließen sich die beiden Argumente nicht zwangsläufig aus bzw. betrachten denselben Prozess von verschiedenen Perspektiven: Das skeptische Argument nimmt die Defizite bestehender Begriffe in den Fokus, das Feintuning-Argument betrachtet eher die Vorteile des neu oder umgestalteten Begriffs (vgl. Koch, 2022, S. 110).

XAI-Kontexten näher fassen will. Etwa in Hinblick auf Fragen wie diese: Unter welchen Bedingungen versteht ein Stakeholder genug, um relevante Desiderata zu erfüllen? Wie wirkt sich das Desideratum auf das jeweils erforderliche Verständnis aus? Müssen unterschiedliche Stakeholder dasselbe Verständnis besitzen, um ein Desideratum erfüllen zu können? Und welche Verstehensermöglicher brauchen wir, um das nötige Verstehen herbeizuführen? Die Beantwortung dieser und vieler anderer Fragen erfordert einen substantiellen Verstehensbegriff, der über allgemeine Verweise auf Verstehen hinausgeht: Wir müssen den Alltagsbegriff für diese spezifischen theoretischen Zwecke feintunen.

4.1.2 Warum Explikation?

Aufgrund der vermehrten philosophischen Auseinandersetzung mit dem Thema kann man als Begriffsgestalterin mittlerweile aus einer Fülle an exemplarischen Vorlagen und sogar schrittweisen ‚Anleitungen‘ wählen (vgl. etwa Brun, 2016; Chalmers, 2020; Isaac et al., 2022). Ich habe mich für ein Verfahren entschieden, das auf Forschung von Gottlob Frege und Bertrand Russell zu Beginn des 20. Jahrhunderts zurückgeht (Greimann, 2007, S. 262): die Explikation.

Der *locus classicus* ist Rudolf Carnaps Behandlung des Themas, insbesondere das erste Kapitel seines *Logical Foundations of Probability*, das erstmals 1950 erschien (Carnap, 1971). Carnap prägte nicht nur den Namen ‚Explikation‘, sondern stellte auch erstmals explizite Regeln für das korrekte Explizieren von Begriffen auf (Greimann, 2007, S. 263).

Die Methode ist in erster Linie durch das in Abschnitt 4.1.1 beschriebene Feintuning-Argument motiviert. Bei der Carnap’schen Explikation handelt es sich um einen begriffsgestaltenden Prozess, bei dem ein unpräziser, präwissenschaftlicher bzw. alltagssprachlicher Begriff (das Explikandum) durch einen mindestens ebenso präzisen neuen Begriff (das Explikat) ersetzt wird, um bestimmte theoretische Zwecke zu erfüllen (vgl. Carnap, 1971, S. 3). Wenn die Explikation erfolgreich ist, besitzt das Explikat gegenüber dem Explikandum bestimmte Vorzüge – etwa in Bezug auf Präzision, Einfachheit oder Fruchtbarkeit –, ohne jedoch eine ausreichende Ähnlichkeit zum vorherigen Begriff einzubüßen.

Carnaps klassisches Beispiel ist die biologische Explikation von *Fisch* in Form von *Piscis* (ebd., S. 5–6): Im vorwissenschaftlichen Sinne bezieht sich ‚Fisch‘ (grob gesagt) auf verschiedene Tiere, die im Wasser leben, z. B. Karpfen und Walfische. ‚Piscis‘ bezeichnet hingegen (in etwa) im Wasser lebende, wechselwarme Wirbeltiere. Das Explikat *Piscis* verändert die Extension des Begriffs *Fisch* – z. B. schließt es Walfische aus – und ist daher bei der biologischen Klassifikation von Tieren von Vorteil (Brun, 2016, S. 1223). Übertragen auf das vorliegende Explikationsprojekt möchte ich in dieser Arbeit ein unpräzises Explikandum, das innerhalb der XAI-Forschung verwendet wird – nämlich den Alltagsbegriff von Verstehen –, durch ein dem Explikandum überlegenes Explikat ersetzen. Der Übersichtlichkeit halber bezeichne ich mein Explikandum, also den Alltagsbegriff von Verstehen, als ‚Verstehen *simpliciter*‘ und mein Explikat als ‚instrumentelles Verstehen‘.

Ich behaupte nicht, dass Carnaps Ansatz die einzige Möglichkeit darstellt, um einen

Begriff instrumentellen Verstehens zu gestalten. Doch scheint mir die Explikation eine besonders attraktive Option darzustellen. Der erste Grund dafür ist, dass Carnaps Methode auf theoretische, wissenschaftliche Vorhaben ausgelegt ist. Zwar ergibt sich die Relevanz meiner Forschungsfragen auch daraus, dass unzureichendes Verstehen in XAI-Kontexten gesellschaftspolitisch und moralisch bedeutsame, negative Konsequenzen haben kann (z. B., wenn der Bias eines KI-Systems unentdeckt bleibt oder wenn durch die Nutzung von KI moralische Verantwortungslücken entstehen). Doch soll mein Verstehensbegriff in erster Linie dabei helfen, den Zusammenhang zwischen Verstehen und Desiderata in XAI-Kontexten theoretisch näher zu fassen und zugehörige (philosophische) Fragestellungen zu beantworten (mehr dazu in Abschnitt 4.2.2).

Damit unterscheidet sich mein Vorhaben deutlich von Begriffsgestaltungen, die vornehmlich politisch, sozial und/oder moralisch motiviert sind. Ein bekanntes Beispiel ist Sally Haslangers ‚ameliorative Analyse‘ von *gender* und *race*. Haslanger definiert diese Begriffe (vereinfacht ausgedrückt) unter Berufung auf systematische Unterdrückung aufgrund wahrgenommener oder zugeschriebener Merkmale (z. B. Geschlechtsorgane oder Hauttöne) mit dem Ziel, dadurch zum Kampf für soziale Gerechtigkeit beizutragen (Haslanger, 2000). Der Gestaltungsprozess von Begriffen, die solchen gesellschaftspolitischen Zielen zuträglich sein sollen, kann sich deutlich von rein theoretisch-akademischen Vorhaben unterscheiden. Beispielsweise ist die Implementierung der ersteren Begriffe meist herausfordernder, weil sie häufiger auf eine ‚globale‘ Veränderung abzielen (vgl. dazu Cappelen, 2018, Kap. 2.3.4). Schließlich müssen nicht nur Wissenschaftler:innen eines Forschungsgebiets von einem Begriffsvorschlag überzeugt werden (was ja schon schwierig genug ist!): Damit z. B. Haslangers Definitionen den gewünschten Effekt entfalten, müssen sie von Aktivist:innen, Politiker:innen oder gar der Breite der Gesellschaft übernommen werden. Außerdem muss ein auf gesellschaftliche Veränderung abzielender Begriff mutmaßlich andere Anforderungen erfüllen als ein rein auf theoretische Kontexte ausgelegter Vorschlag – etwa hinsichtlich der Einfachheit und Verständlichkeit für Nichtfachleute. Für meine Zwecke ist daher eine Methode attraktiv, die auf theoretisch-wissenschaftliche Vorhaben zugeschnitten ist.

Es gibt noch einen zweiten Grund, weswegen ich mich für die Explikation entschieden habe: Die Methode ist theoretisch vergleichsweise gut entwickelt und bietet ein konkretes Schema für den Prozess der Begriffsgestaltung. Im Zuge dieser Weiterentwicklung wurde die Methode zudem von theoretischen Annahmen gelöst, die spezifisch für Carnap (oder die Doktrinen des Wiener Kreises) sind.⁴² Dadurch besteht keine Notwendigkeit, exegetisch bzw. philosophiehistorisch zu arbeiten, wenn man eine Explikation durchführen will.

Ich bin im Übrigen nicht die Erste, die eine an Carnap angelehnte Explikation durchführt, um Verstehen zu explizieren: Christoph Baumberger und Georg Brun explizieren anhand dieser Methode objektuales Verstehen (Baumberger & Brun, 2017), einen ähn-

⁴²Beispielsweise identifiziert Catarina Dutilh Novaes Parallelen zwischen Haslangers ameliorativer Analyse und Carnaps Explikation (Dutilh Novaes, 2020).

lichen Vorschlag formuliert Baumberger auch in einem späteren Aufsatz (Baumberger, 2019). Diese Aufsätze sind ein wichtiges Vorbild für meine Definition instrumentellen Verstehens in Kapitel 11, wenngleich mein Vorschlag von den dort vertretenen Positionen abweicht. Stefan Sleeuw expliziert in seiner Dissertation symbolisches und faktisches objektuales Verstehen (Sleeuw, 2023). Anders als ich distanziert er sich radikal von der gängigen erkenntnistheoretischen Debatte und entwickelt eine behaviorale Definition dieser beiden Verstehensformen. In ähnlicher Weise expliziert Mark Newman wissenschaftliches Verstehen einer Theorie durch die Fähigkeit, Schlüsse zu ziehen, die in Lösungen für qualitative Probleme resultieren, die durch diese Theorie adressiert werden können (Newman, 2017b). Meinem Vorschlag zufolge bilden Fähigkeiten hingegen nur eine von mehreren Bedingungen instrumentellen Verstehens (siehe Kapitel 9).

4.1.3 Das Rezept

Ich strukturiere meine Explikation gemäß des ‚Rezepts fürs Explizieren‘, das Georg Brun formuliert hat (Brun, 2016, Kap. 3). Eine Explikation erfolgt demzufolge in vier Schritten:

1. Die *Spezifikation der Explikationsaufgabe* beinhaltet
 - (a) die Präzisierung des Explikandums,
 - (b) die Spezifikation des theoretischen Rahmens, in den das Explikat eingebettet werden soll und
 - (c) die Benennung der Funktion(en) bzw. Zwecke, die das Explikat in diesem Rahmen erfüllen soll.
2. Bei der *Spezifikation von Adäquatheitsbedingungen* werden
 - (a) Ähnlichkeits- und
 - (b) Nützlichkeitsbedingungenformuliert, die das Explikat erfüllen muss.
3. Die *Einführung des Explikats* besteht in der Formulierung von Verwendungsregeln für das Explikat innerhalb der Zieltheorie (gängigerweise in Form einer Definition).
4. Im letzten Schritt findet die *Evaluation des Explikats* anhand der im zweiten Schritt formulierten Adäquatheitskriterien statt.

Diese vier Schritte gliedern mein Explikationsprojekt. Zur Erinnerung: Das Explikandum ist der Alltagsbegriff von Verstehen, den ich als ‚Verstehen *simpliciter*‘ bezeichne. Das Explikat ist ein Begriff jener Form des Verstehens, welche die Erfüllung von Desiderata in XAI-Kontexten ermöglicht. Ich nenne diese Verstehensform ‚instrumentelles Verstehen‘. Im verbleibenden Teil dieses Kapitels widme ich mich dem ersten und zweiten Schritt des Rezepts, also der Spezifikation der Explikationsaufgabe (Abschnitt 4.2) und der Adäquatheitsbedingungen für das Explikat, d. h. für *instrumentelles Verstehen* (Abschnitt 4.3). Um

die Erfüllung dieser Adäquatheitsbedingungen zu gewährleisten, setze ich mich in den Kapiteln 5 bis 10 intensiv mit philosophischen Debatten zum Thema Verstehen auseinander. Auf dieser Grundlage erfolgt der dritte Schritt, also die Einführung des Explikats, in Kapitel 11. In demselben Kapitel schließe ich meine Explikation mit dem vierten Schritt ab, nämlich der Evaluation des Explikats anhand der im zweiten Schritt formulierten Adäquatheitsbedingungen.

4.2 Die Explikationsaufgabe

In diesem Abschnitt widme ich mich dem ersten Schritt des in Abschnitt 4.1.3 beschriebenen Rezepts: der Konkretisierung der Explikationsaufgabe. Dieser Schritt ist in drei Teilschritte zerlegt. Zunächst charakterisiere ich das Explikandum, indem ich für mein Vorhaben relevante Bedeutungen von Verstehen von anderen Bedeutungen abgrenze und die hier maßgeblichen Verstehenssubjekte bestimme (Abschnitt 4.2.1). Anschließend lege ich den theoretischen Rahmen fest, in den sich instrumentelles Verstehen einfügen soll (Abschnitt 4.2.2). Zum Schluss konkretisiere ich die theoretischen Zwecke, die das Explikat in diesem Rahmen erfüllen soll (Abschnitt 4.2.3).

4.2.1 Präzisierung des Explikandums

Der erste Schritt (1a) des Rezepts besteht in der Präzisierung des Explikandums. Denn um einen Begriff explizieren zu können, muss man diesen Begriff zunächst bestimmen und deutlich machen, was man damit meint (und was nicht). Im vorliegenden Fall genügt es also nicht, zu sagen, dass ich *Verstehen* explizieren möchte. Denn Verstehen umfasst eine Vielzahl an kognitiven Zuständen, Vermögen, Prozessen und Errungenschaften, von denen nur ein Bruchteil für XAI-Kontexte relevant ist. Zudem kommen grundsätzlich eine Reihe verschiedener Verstehenssubjekte für meine Explikation in Frage. *Verstehen* ist in diesem Sinne unpräzise (daher auch der Bedarf der Begriffsgestaltung) und kann zu Beginn der Explikation entsprechend auch nur unpräzise beschrieben werden. Dennoch muss man sich, so Carnap, darum bemühen, durch Beispiele und informelle Definitionen möglichst klar zu machen, was mit dem Explikandum gemeint ist (Carnap, 1971, §2). Das ist das Ziel dieses Abschnitts.

Relevante Bedeutungen von ‚Verstehen‘

Manche Bedeutungen von ‚Verstehen‘ scheiden bereits deswegen als Explikandum aus, weil sie aus erkenntnistheoretischer Perspektive uninteressant sind: Beispielsweise drücken wir einem Freund unser Verständnis für seine missliche Lage aus; im Englischen bedeutet ‚understanding‘ auch *Mitgefühl*; und der Satz ‚Ich verstehe, dass du wütend bist‘, soll in erster Linie beschwichtigen (vgl. dazu Elgin, 2017, S. 38). Instrumentelles Verstehen entspricht hingegen eher jenen Verwendungsweisen, die einen kognitiven oder epistemischen

Erfolg betreffen – zum Beispiel den epistemischen Erfolg, ein KI-System zu verstehen.

Instrumentelles Verstehen entspricht außerdem nicht dem *Gefühl* von Verstehen, also so etwas wie dem berühmten *Heureka!*-Moment, der Archimedes zu einem unbekleideten Spaziergang von einer Badeanstalt nach Hause veranlasst haben soll. Zwar kennen wir alle das Aha-Gefühl, wenn es endlich ‚Klick‘ macht. Doch dieses Gefühl ist weder notwendig noch hinreichend für Verstehen – so mancher Aha-Moment beruht auf einem Missverständnis, und oftmals bemerken wir gar nicht, dass wir etwas verstanden haben.⁴³ Leonid Rozenblit und Frank Keil haben diese theoretische Annahme aus psychologisch-kognitionswissenschaftlicher Sicht (unter dem Schlagwort ‚*illusion of explanatory depth*‘) empirisch bestätigt: Menschen neigen dazu, ihr eigenes (explanatorisches) Verstehen zu überschätzen (z. B. Rozenblit & Keil, 2002). Das trügerische Gefühl von Verstehen ist also nicht das Explikandum, um das es gehen soll.

Andere Bedeutungen von ‚Verstehen‘ kommen zwar als Gegenstand erkenntnistheoretischer Untersuchungen in Frage, sind für meine Zwecke aber uninteressant. Nelson Goodman und Catherine Elgin unterscheiden zwischen Verstehen als (1) Fähigkeit oder Vermögen, (2) Prozess und (3) Ergebnis eines Prozesses (N. Goodman & Elgin, 1993, S. 213; vgl. auch Baumberger, 2011, S. 69–70; M. P. Lynch, 2018). Im Sinne von (1) besteht Verstehen in verschiedenen (kognitiven) Fähigkeiten, wie forschen, erfinden, unterscheiden, identifizieren, organisieren, prüfen oder verwerfen. Verstehen als (2) Prozess meint die Anwendung oder Verbesserung dieser Fähigkeiten, um Verstehen zu vertiefen oder zu erweitern. Drittens ist Verstehen auch (3) das Ergebnis der Anwendung kognitiver Fähigkeiten, ein (revidierbarer) epistemischer Zustand (N. Goodman & Elgin, 1993, S. 213). Dieser Zustand kann ebenfalls durch spezifische Fähigkeiten gekennzeichnet sein (siehe Kapitel 9). Instrumentelles Verstehen ist dem letztgenannten Sinn am nächsten: Es handelt sich um den epistemischen Zustand, der die Erfüllung gesellschaftlicher Desiderata in XAI-Kontexten ermöglicht.⁴⁴

Relevante Verstehenssubjekte

Wer kommt als Subjekt instrumentellen Verstehens in Frage? Grundsätzlich ist festzuhalten, dass mein Explikat ausschließlich das Verstehen von Personen einschließt. Zwar werfen die wachsenden Fähigkeiten moderner KI-Systeme die Frage auf, unter welchen Bedingungen

⁴³Vgl. dazu etwa Baumberger, 2011, S. 73; Delarivière & Van Kerkhove, 2021, S. 626; de Regt, 2009, S. 26; Kuorikoski, 2012; P. Lipton, 2009, S. 57; Trout, 2002, 2007; Ylikoski, 2009. Auch wenn er die Unzuverlässigkeit des ‚Aha-Gefühls‘ anerkennt, denkt Peter Lipton darüber nach, wie dieses Gefühl die Entscheidung von Wissenschaftler:innen für eine Erklärung oder Theorie anleiten könnte (P. Lipton, 2009, S. 54–60). Die Frage nach der Notwendigkeit des Aha-Gefühls hängt auch mit der Debatte zur Transparenz von Verstehen zusammen (siehe Fußnote 32).

⁴⁴Das heißt nicht, dass (1) und (2) in XAI-Kontexten keine Rolle spielen. Eva Schmidt hat mich richtigerweise darauf hingewiesen, dass Erklärbarkeit auch Verstehen als Vermögen und Prozess zuträglich sein kann. Das ergibt sich bereits daraus, dass Verstehen als Zustand das Ergebnis von Vermögen und Prozess des Verstehens ist.

man diesen Systemen Verstehen zuschreiben muss (vgl. Barman et al., 2024; M. Mitchell & Krakauer, 2023; Stuart, 2025; Thórisson et al., 2016; Wilkenfeld, 2013). Allerdings folge ich in dieser Arbeit der Auffassung, dass KI-Systeme Werkzeuge zur Lösung spezifischer Aufgaben sind (siehe Kapitel 2). Unterschiedliche Mitglieder der Gesellschaft tragen verschiedene Desiderata an die Entwicklung, Implementierung und Verwendung dieser Systeme heran, und instrumentelles Verstehen soll zur Erfüllung solcher Desiderata beitragen. Das Subjekt instrumentellen Verstehens ist daher stets ein Mensch, und KI-Systeme kommen lediglich als Objekte dieses Verstehens in Frage.

Wie viele andere Autor:innen in der XAI-Debatte bezeichne ich die unterschiedlichen Personen, die als Verstehenssubjekte in Frage kommen, als ‚Stakeholder‘. Das Merriam-Webster-Wörterbuch definiert einen Stakeholder sehr breit als „one who is involved in or affected by a course of action“. Übertragen auf XAI kommen dieser Auffassung nach prinzipiell alle Personen als Stakeholder in Frage, auf die sich ein konkretes KI-System in relevantem Maße auswirkt oder die anderweitig in Nutzung oder Entwicklung des Systems eingebunden sind.

Harini Suresh et al. identifizieren zwei Strategien, die häufig gewählt werden, um verschiedene Stakeholder in XAI-Kontexten voneinander zu unterscheiden: Einerseits rollenbasierte, andererseits expertisebasierte Stakeholdergruppen (Suresh et al., 2021). *Rollenbasierte* Stakeholdergruppen sind beispielsweise die Entwickler:innen eines KI-Systems; von den Entscheidungen des KI-Systems betroffene Gruppen oder Einzelpersonen; ausführende Personen, die KI-gestützte Entscheidungen umsetzen; Menschen in der Schleife, die KI-Empfehlungen vor der Umsetzung prüfen; ethische oder legislative Normengeber:innen; Prüfer:innen, die das KI-System auditieren; oder auch Datensubjekte, deren Daten in die Erstellung des Systems eingeflossen sind (siehe z. B. Langer et al., 2021; Tomsett et al., 2018). Mögliche *expertisebasierte* Stakeholdergruppen sind KI-Forscher:innen, Fachexpert:innen mit Expertise in der Anwendungsdomäne des Systems oder Endnutzer:innen bzw. Nichtfachleute (siehe etwa Ribera & Lapedriza, 2019; Yu & Shi, 2018). Darüber hinaus gibt es mehrere Stakeholder-Taxonomien, die Rollen- und Expertisebezeichnungen kombinieren (z. B. Gerlings et al., 2022; Preece et al., 2018). Diese beiden Perspektiven auf Stakeholder – Rolle und Expertise – korrelieren offensichtlich miteinander. Die Entwicklerin eines KI-Systems (Rolle) kennt sich üblicherweise mit informatischen Aspekten dieses Systems aus (Expertise); ein Mensch in der Schleife (Rolle) besitzt normalerweise Fachwissen hinsichtlich der Anwendungsdomäne des Systems (Expertise) usw.

Im Kontext dieser Arbeit ist der folgende *normative* Zusammenhang zwischen der Rolle und Expertise von Stakeholdern wichtig: Wir erwarten von der Entwicklerin, dass sie informatische Expertise besitzt bzw. von dem Menschen in der Schleife, dass er sich in der Anwendungsdomäne des Systems auskennt, und diese Erwartungen sind aufgrund der jeweiligen Rolle angemessen.⁴⁵ Ähnliche Überlegungen finden sich in Debatten zu

⁴⁵Diesen Gedanken übernehme ich von Adam Carter und Emma Gordon. Die Autor:innen schlagen vor, dass der Besitz von Verstehen manchmal eine relevante epistemische Norm des Behauptens darstellt: Um

‚vorwärtsgerichteter‘, also prospektiver Verantwortung. Herbert Hart schlägt vor, dass aus der Rolle einer Person spezifische Verantwortlichkeiten bzw. Pflichten erwachsen (engl. *role responsibilities*). Eine Kapitänin ist beispielsweise *qua* ihrer Rolle für die Sicherheit des Schiffs verantwortlich (Hart, 2008, S. 212–214). Analog könnte man behaupten, dass z. B. eine Systementwicklerin aufgrund ihrer Rolle für die ausreichende Zuverlässigkeit und Sicherheit eines KI-Systems verantwortlich ist. Die Erfüllung dieser Pflichten kann wiederum erfordern, dass sie relevante technische Expertise besitzt. Andere Stakeholder (z. B. Entscheidungsobjekte) haben diese Pflichten nicht. Es wäre daher auch nicht angemessen, von ihnen die Expertise zu verlangen, die für die Erfüllung dieser Pflichten nötig ist. So oder so: Es ist plausibel anzunehmen, dass die spezifische Rolle eines Stakeholders legitime Erwartungen hinsichtlich Form und Grad seiner Expertise weckt.

Expertise lässt sich als eine Form des Verstehens auffassen, wenn sie bestimmte mentale Zustände (z. B. wahre Überzeugung) und Fähigkeiten beinhaltet oder erfordert. Eine Expertin bezüglich einer Domäne *D* hat z. B. mehr wahre Überzeugungen hinsichtlich *D* als ein Laie und kann Aufgaben und Probleme in *D* lösen, die der Laie nicht bewältigen könnte (ggf. existiert ein Schwellwert, der für ein Mindestmaß an wirklicher Expertise überschritten werden muss).⁴⁶ Diese beiden Elemente von Expertise – geeignete kognitive Zustände hinsichtlich domänenspezifischer Inhalte sowie das Verfügen über relevante Fähigkeiten – finden sich in der Überzeugungs- und Fähigkeitskomponente von Verstehen wieder, die ich in Kapitel 7 bzw. Kapitel 9 ausführlich erläutern werde.

Der normative Zusammenhang zwischen Rolle und Expertise wird im weiteren Verlauf der Arbeit dabei helfen, das instrumentelle Verstehen eines Stakeholders in einem bestimmten Kontext zu bestimmen (siehe Abschnitte 8.4 und 9.3 und Kapitel 11). Denn die Erfüllung legitimer Expertiseerwartungen ist eine wichtige Voraussetzung für die erfolgreiche epistemische Arbeitsteilung zwischen Stakeholdern. In Abschnitt 2.3.3 wurde deutlich, dass diese Arbeitsteilung nicht nur eine Quelle von epistemischer Opazität, sondern auch eine Strategie zu deren Überwindung sein kann. Mein Begriff instrumentellen Verstehens macht sich diesen Vorteil zunutze, indem er die normative Dimension gelungener epistemischer Arbeitsteilung heranzieht, um instrumentelles Verstehen für Stakeholder mit unterschiedlichster Expertise erreichbar zu machen. Mein Vorschlag wird *nicht* lauten, dass ein gegebener Stakeholder aufgrund seiner Rolle eine absolute Verpflichtung besitzt, instrumentell zu verstehen. Es ist durchaus möglich, dass die vielen, teils schwer zu über-

korrekterweise zu behaupten, dass *p*, muss man mitunter explanatorisch verstehen, warum *p* als wahr beurteilt wird (Carter & Gordon, 2011). Das Innehaben einer Rolle durch die behauptende Person weckt in diesem Fall legitime Erwartungen hinsichtlich ihrer Expertise (hier: eine bestimmte Form explanatorischen Verstehens) und lässt die beschriebene Norm des Behauptens greifen (ebd.; Gordon, 2021, Kap. 5). Nebenbei bemerkt: Diese Überlegungen legen nahe, dass die Verwendung von opaken KI-Systemen das Einhalten relevanter Normen des Behauptens verhindern kann (vgl. auch Schmidt et al., 2025, Fn. 11 auf S. 7). Wenn angemessenes Behaupten mitunter Verstehen erfordert, und opake KI-Systeme dieses Verstehen verhindern, wäre dies ein Argument für die Notwendigkeit von erklärbarer KI bzw. gegen die Nutzung opaker Systeme.

⁴⁶Diese Darstellung stützt sich auf die verschiedenen Definitionsvorschläge für Expertise von Alvin Goldman (2018, Kap. 1).

windenden Quellen von Opazität (siehe Abschnitt 2.3) (instrumentelles) Verstehen von KI für einen gegebenen Stakeholder unmöglich machen – auch dann, wenn er jegliche Expertise besitzt, die wir *qua* seiner Rolle legitimerweise von ihm erwarten können. Die Idee ist vielmehr, dass sich aus Stakeholder-Rollen eine legitime epistemische Arbeitsteilung in XAI-Kontexten ableiten lässt. Und diese Arbeitsteilung kann als Heuristik dienen, um instrumentelles Verstehen in einem kontextspezifischen Sinne zu charakterisieren.

Zusammengefasst: Mein Explikandum ist ein epistemischer Zustand, der Stakeholdern mit unterschiedlichen Rollen und Fachkenntnissen die Erfüllung gesellschaftlicher Desiderata in XAI-Kontexten ermöglicht.

4.2.2 Der theoretische Rahmen des Explikats

Kommen wir zum nächsten Schritt (1b) des Explikations-Rezepts. Bei begriffsgestaltenden Vorhaben ist es wichtig, den Geltungsbereich bzw. theoretischen, gesellschaftlichen oder sonstigen Rahmen zu bestimmen, in den der neue bzw. veränderte Begriff eingegliedert werden soll. Beispielsweise zielt das Explikat *Piscis* nicht darauf ab, den Begriff *Fisch* in allen Kontexten zu ersetzen, sondern lediglich im theoretischen Rahmen der Zoologie.

Mein Vorhaben soll sich in erster Linie in erkenntnistheoretische Arbeiten einfügen, insofern sie sich mit dem Zusammenhang zwischen Erklärbarkeit, Verstehen und gesellschaftlichen Desiderata auseinandersetzen. Zwar stütze ich mich auch auf wissenschaftstheoretische Arbeiten zum Thema Verstehen, doch nimmt meine Explikation unter anderem Bezug auf Debatten zu Überzeugung, Rechtfertigung und Faktizität, die vor allem innerhalb der Erkenntnistheorie verhandelt werden. Nichtsdestotrotz sollte *instrumentelles Verstehen* auch für Wissenschaftstheoretiker:innen anschlussfähig sein.

Aufgrund der Interdisziplinarität der Erklärbarkeitsforschung möchte ich zusätzlich noch Theoriesysteme aus anderen Disziplinen in den erweiterten Kreis der angepeilten theoretischen Rahmen aufnehmen. Erklärbarkeit ist ein interdisziplinäres Unterfangen, das unter anderem informatische, psychologische, juristische und philosophische Forschung erfordert (Langer et al., 2021). Idealerweise würden diese verschiedenen Perspektiven zu einem gemeinsamen Forschungsprogramm gebündelt. Doch das setzt voraus, dass innerhalb der einzelnen Disziplinen Klarheit herrscht. Mein Vorschlag soll auch zu diesem größeren Projekt beitragen, indem es zur Klärung von Verstehen aus philosophischer Perspektive beiträgt.

Die Zielgruppe meines Vorschlags sind also hauptsächlich Philosoph:innen, die sich für den Zusammenhang zwischen Verstehen und Desiderata in XAI-Kontexten interessieren. Anders als z. B. bei Sally Haslanger, die mit ihrer ameliorativen Analyse von *gender* und *race* auch eine gesellschaftspolitische Agenda verfolgt, richtet sich mein Begriff instrumentellen Verstehens demnach auf einen theoretischen Kontext. Entsprechend ist es nicht mein Ziel, meinen Begriffsvorschlag in der Alltagssprache zu implementieren. Allerdings könnte eine erfolgreiche Implementierung in der Fachgemeinschaft auch gesellschaftlich vorteilhaft sein: Wenn die XAI-Forschung einen substanziellen Verstehensbegriff zur Verfügung hat,

könnte dies die Entwicklung von geeigneten Verstehensmöglichkeiten befördern (mehr dazu in Kapitel 12). Und dies wäre bestenfalls der Erfüllung relevanter Desiderata zuträglich.

4.2.3 Die theoretischen Zwecke des Explikats

Ich widme mich nun Schritt (1c) des Rezepts und erläutere, welche theoretischen Funktionen bzw. Zwecke⁴⁷ dem Explikat innerhalb des angepeilten theoretischen Rahmens (siehe Abschnitt 4.2.2) zukommen sollen.

Zur Klarstellung: Damit sind nicht die gesellschaftlichen Desiderata gemeint, die mithilfe instrumentellen Verstehens erfüllt werden sollen. Auch sind die Zwecke des Explikats nicht identisch mit den Adäquatheitsbedingungen, die ich in Abschnitt 4.3 aufstelle – wenngleich die Erfüllung der theoretischen Zwecke eines der Adäquatheitskriterien darstellt. Es geht um die Funktionen bzw. Zwecke des *Begriffs* von instrumentellem Verstehen. Doch was ist der Zweck eines Begriffs?⁴⁸ Um sich dieser Frage in Bezug auf mein Explikat anzunähern, ist es hilfreich, zunächst das *Phänomen* zu betrachten, das dadurch eingefangen werden soll.

In Kapitel 1 und Abschnitt 4.1.1 habe ich im Anschluss an Markus Langer et al. (ebd.) herausgearbeitet, dass Verstehen in der Erklärbarkeitsliteratur als Mittel zur Erfüllung gesellschaftlicher Desiderata angesehen wird. Dieser vielfach geäußerten Intuition zufolge handelt es sich hierbei also um einen wichtigen Zweck von Verstehen in XAI-Kontexten. Doch welche Desiderata sind das? Und wie hängen sie mit Verstehen zusammen? Ich beleuchte zunächst diese Fragen, um auf dieser Grundlage die theoretischen Zwecke des Explikats zu identifizieren.

Mögliche Desiderata von Stakeholdern

Stakeholder können prinzipiell eine Vielzahl sehr unterschiedlicher Desiderata besitzen. Beispielsweise sollten medizinische Diagnostik-Systeme *vertrauenswürdig* sein, ein Empfehlungssystem für Bewerber:innen muss *fair* sein, und der Algorithmus eines Musikstreamingdienstes sollte möglichst *genau* sein und Lieder vorschlagen, die dem Geschmack der Nutzer:innen entsprechen. Im Rahmen einer ausführlichen Literaturschau haben Markus Langer et al. 29 verschiedene Desiderata identifiziert, die in der Literatur diskutiert werden (ebd., Tab. 1 auf S. 5):

⁴⁷Viele Begriffsgestalter:innen verwenden ‚Funktion‘ im Kontext von Begriffen synonym zu ‚Zweck‘ oder auch ‚Ziel‘. Ich folge dieser Konvention und meine nicht den anspruchsvollen Funktionsbegriff, der beispielsweise in der Debatte zur Funktion von Artefakten oder organischen Strukturen genutzt wird.

⁴⁸Diese Frage wird in der Literatur zur Methodik der Begriffsgestaltung kontrovers diskutiert; vgl. etwa Cappelen, 2018, Kap. 16; Nado, 2021; Simion & Kelp, 2020; Thomasson, 2024. Beispielsweise argumentiert Herman Cappelen, dass die einzige Funktion eines Begriffs *B* darin besteht, *B*s zu bezeichnen bzw. ihre Denotata zu beschreiben (Cappelen, 2018, Kap. 16). Sollte Cappelen Recht haben, müssten die folgenden Überlegungen eher als weitere Präzisierung des Explikandums aufgefasst werden.

- Akzeptanz des Systems
- Fähigkeit, jemanden bei Fehlern zur Rechenschaft zu ziehen (engl. *accountability*)
- Genauigkeit des Systems
- Wahrung der menschlichen Autonomie
- Zutrauen der Nutzer:innen ins System (engl. *confidence*)
- Wahrung der menschlichen Kontrolle
- Beseitigbarkeit von Fehlern
- Erlernen des Umgangs mit einem System (engl. *education*)
- Effizienz des Systems
- Effektivität
- Fairness
- informierte Einwilligung in die Entscheidungen eines Systems
- Gesetzeskonformität
- Übereinstimmung des Systems mit moralischen und ethischen Standards
- Leistungsfähigkeit
- Datenschutz
- Wahrung der menschlichen Verantwortlichkeit (engl. *responsibility*)
- Robustheit
- Sicherheit (sowohl im Sinne von *safety* als auch *security*)⁴⁹
- Zufriedenheit mit dem System
- Erlangung wissenschaftlicher Einsichten durch das System
- Möglichkeit zum Transfer des Modells in andere Kontexte
- Transparenz
- Vertrauen
- Vertrauenswürdigkeit
- Nutzbarkeit
- Nützlichkeit
- Verifikation

Diese Auflistung legt nahe, dass in XAI-Kontexten mindestens vier Gruppen von Desiderata relevant sind: erstens Desiderata, die eine wünschenswerte Eigenschaft von KI-Systemen darstellen (z. B. Fairness, Vertrauenswürdigkeit, Genauigkeit); zweitens Desiderata, die in wünschenswerten Eigenschaften des umliegenden soziotechnischen Systems bestehen (etwa die Möglichkeit, jemanden bei Fehlern zur Rechenschaft zu ziehen); drittens Desiderata, welche die Interaktion oder Arbeit mit einem KI-System innerhalb des soziotechnischen Systems betreffen (beispielsweise die Wahrung der menschlichen Verantwortlichkeit oder Autonomie). Viertens bestehen manche Desiderata in einer wünschenswerten Einstellung eines Menschen gegenüber dem System (ebenfalls im Rahmen des übergeordneten soziotechnischen Systems), z. B. Vertrauen oder Akzeptanz.

Natürlich kann diese Liste keinen Anspruch auf Vollständigkeit erheben, doch fasst sie häufig diskutierte Desiderata zusammen und illustriert die Vielfalt der möglichen Anforderungen an KI-Systeme oder Kontexte, in denen sie implementiert werden. Es gibt allerdings eine wichtige Einschränkung: In dieser Arbeit bleiben moralisch fragwürdige oder eindeutig verwerfliche Desiderata unberücksichtigt, wie die Steigerung des Profits von KI-Unternehmen, die Manipulation von Endnutzer:innen, das illegale Auslesen perso-

⁴⁹ *Safety* beschreibt im Englischen die Verhinderung von Unfällen (z. B. durch Technologie zur automatischen Abschaltung des Systems in Gefahrensituationen), während *security* die Verhinderung von Straftaten umfasst (z. B. durch Maßnahmen zum Datenschutz). Im Deutschen gibt es für diese beiden Facetten von Sicherheit keine eigenen Ausdrücke, daher liste ich sie nicht einzeln auf.

nenbezogener Daten oder das Verschleiern von Fehlern oder Biases durch irreführende Erklärungen (siehe dazu Bordt et al., 2022). Zudem schließe ich unangemessene Einstellungen gegenüber KI-Systemen (z. B. zu geringes Vertrauen oder übermäßige Akzeptanz und Zufriedenheit) aus. In XAI-Kontexten sollte das Ziel sein, eine angemessene Einstellung gegenüber dem fraglichen KI-System einzunehmen. Beispielsweise sollte der hohe bzw. niedrige Vertrauensgrad einer Anwenderin in ein KI-System dessen tatsächlicher hoher bzw. geringfügiger Vertrauenswürdigkeit entsprechen (Schlicker et al., 2025). Mit ‚(gesellschaftlichen) Desiderata‘ meine ich in dieser Arbeit ausschließlich solche Anforderungen an KI-Systeme bzw. KI-Kontexte, die in dieser Form angemessen bzw. moralisch vertretbar sind.

Weiterhin ist nicht jedes Desideratum für alle Stakeholder gleichermaßen relevant (Langer et al., 2021, Kap. 2.2): Für KI-Entwickler:innen stehen Desiderata wie Effizienz oder Beseitigbarkeit von Fehlern stärker im Vordergrund als für Endnutzer:innen; Fairness und Übereinstimmung mit anderen moralischen oder ethischen Standards spielt eine besonders wichtige Rolle für Personen, die von den Entscheidungen eines Systems betroffen sind; das Datenschutz-Desideratum ist insbesondere für Datensubjekte wichtig, deren Daten in das System einfließen, usw.

Verstehen als Mediator zwischen Erklärbarkeit und Desiderata

Die Annahme, dass Verstehen die Brücke zur Erfüllung von Desiderata in XAI-Kontexten schlägt, ist zwar weit verbreitet und intuitiv plausibel. Meist bleibt sie aber implizit. Markus Langer et al. haben ein begriffliches Modell entwickelt, das diesen Zusammenhang explizit macht und Verstehen als *Mediator* zwischen Erklärbarkeit und gesellschaftlichen Desiderata auffasst (ebd.). Die Autor:innen treffen eine hilfreiche Unterscheidung zwischen der epistemischen und substanziellen Erfüllung eines Desideratums (ebd., Kap. 3.1). Beispielsweise ist das Desideratum Fairness *substanziell* erfüllt, wenn das fragliche KI-System tatsächlich im relevanten Sinne fair ist. *Epistemisch* ist das Desideratum hingegen erfüllt, wenn der fragliche Stakeholder in der Lage ist zu wissen oder zu beurteilen, ob das System fair ist oder nicht bzw. ob es mit Fairness korrespondierende Eigenschaften besitzt. Entsprechend kann ein Desideratum substanziell erfüllt sein, ohne (für einen bestimmten Stakeholder) epistemisch erfüllt zu sein, und umgekehrt.

Meist ist das Verhältnis zwischen epistemischer und substanzieller Erfüllung von Desiderata aber subtiler. Beispielsweise sind manche Desiderata zugleich substanziell erfüllt, sobald sie epistemisch erfüllt sind (ebd., S. 8). Ein Beispiel ist das Desideratum der menschlichen Beaufsichtigung (engl. *human oversight*), also die Überwachung und ggf. Kontrolle von KI-Systemen und KI-gestützten Entscheidungen durch einen Menschen: Wenn man ein KI-System zu einem hohen Grad versteht, ist die menschliche Beaufsichtigung auch in substanzieller Hinsicht gegeben. Ein anderes Beispiel ist die informierte Einwilligung in KI-gestützte Entscheidungen – sobald man relevante systembezogene Aspekte hinreichend versteht, ist dieses Desideratum auch substanziell erfüllt.

In anderen Fällen macht die epistemische Erfüllung eines Desideratums zumindest einen Teil seiner substanziellen Erfüllung aus. Beispielsweise wird moralische Verantwortlichkeit für Handlungen üblicherweise durch zwei notwendige und zusammen hinreichende Bedingungen bestimmt: Die Kontrollbedingung verlangt (vereinfacht ausgedrückt), dass die fragliche Person hätte anders handeln können bzw. aus freien Stücken gehandelt hat. Die epistemische Bedingung fordert, dass der handelnden Person bewusst war, was sie getan hat – z. B. in Hinblick auf die moralisch relevanten Konsequenzen ihrer Handlung. Mithilfe geeigneter Verstehensermöglicher kann ein Stakeholder diese epistemische Bedingung bei KI-gestützten Entscheidungen erfüllen und, sofern auch die Kontrollbedingung erfüllt ist, (substanziell) moralisch verantwortlich für seine Handlung sein (siehe dazu Baum et al., 2022). Kurz: Bei Desiderata, die epistemischer Natur sind oder zumindest eine epistemische Komponente haben, kann Erklärbarkeit sowohl zur epistemischen als auch substanziellen Erfüllung beitragen, weil beide (wenigstens teilweise) zusammenfallen.

Die meisten Desiderata können hingegen bestenfalls epistemisch, aber nicht substanziell durch Transparenz von KI-Systemen erfüllt werden. Zum Beispiel setzt die substanzielle Erfüllung des Desideratums, jemanden bei Fehlern zur Rechenschaft zu ziehen, auch einen geeigneten institutionellen Rahmen voraus. Allerdings kann die epistemische Erfüllung eines Desideratums den Ausgangspunkt für dessen substanzielle Erfüllung bieten. Erklärbarkeit könnte zum Beispiel aufdecken, dass ein System unfair ist – das Desideratum wäre dann in epistemischer Hinsicht erfüllt. Ausgehend von dieser Einsicht können Entwickler:innen das System dann verbessern und das Fairness-Desideratum auch substanziell erfüllen (Langer et al., 2021, S. 8).

Interessanterweise verhindert die epistemische Erfüllung eines Desideratums mitunter auch dessen substanzielle Erfüllung. Beispielsweise verwenden die Polizeibehörden mehrerer deutscher Bundesländer eine Variante der Data-Mining-Software *Gotham* zur präventiven Kriminalitäts- und Terrorbekämpfung. Diese Software stammt vom US-Unternehmen *Palantir*, das vom bekennenden Antidemokraten Peter Thiel mitgegründet wurde. *Gotham* führt personenbezogene Daten aus diversen Quellen zusammen und wertet sie aus. Dass Polizeibeamt:innen – und schlimmstenfalls sogar Mitarbeiter:innen von *Palantir* – durch *Gotham* Einblicke in massive Mengen sensibler Daten erhalten können, wird von Datenschützer:innen seit Jahren kritisiert (Kurz & Ullrich, 2025). Ein ausgeprägtes Verständnis der Analysesoftware und ihrer Vorhersagen gefährdet offenkundig den substanziellen Datenschutz derjenigen, deren Daten gesammelt und ausgewertet werden. Hier kommt es auch darauf an, welcher Stakeholder Verstehen besitzt: Ein Verständnis der Verarbeitung meiner eigenen persönlichen Daten gefährdet die substanzielle Erfüllung meines Datenschutz-Desideratums nicht. Das ist anders, wenn ein Polizeibeamter mein Privatleben durchleuchtet.

Schließlich wird die epistemische Facette eines Desideratums dadurch beeinflusst, ob die Verbesserung des Systems für einen Stakeholder im Fokus steht. Nehmen wir eine Systementwicklerin, die ein KI-System z. B. substanziell fair, sicher oder robust machen will und deswegen XAI-Ansätze nutzt, um das System besser zu verstehen. Diese Entwicklerin

braucht ein Verständnis, das sie in die Lage versetzt, das Desideratum substanziell zu erfüllen. Beispielsweise benötigt sie ein Verständnis der Ursachen etwaiger Probleme und der Mittel, um diese zu beheben. Eine typische Endnutzerin, die im Berufsalltag mit dem System arbeitet, will hingegen bloß herausfinden, *ob* das System fair, sicher oder robust ist – sie hat nicht den Anspruch, etwaige Probleme zu beseitigen. Wenn sie z. B. herausfindet, dass das System unfair ist, ist ihr Desideratum epistemisch erfüllt, auch wenn sie keine Ahnung hat, wie sie den Bias beseitigen kann.⁵⁰

Die epistemische und substanzielle Erfüllung von Desiderata in XAI-Kontexten interagiert also in komplexer Weise miteinander. Für Erklärbarkeit und das dadurch hervorbrachte Verstehen ist vornehmlich die epistemische Variante entscheidend. Daher richtet sich auch instrumentelles Verstehen üblicherweise auf die epistemische Erfüllung gesellschaftlicher Desiderata. Oben wurde allerdings deutlich, dass die epistemische und die substanzielle Erfüllung bei manchen Desiderata wenigstens teilweise zusammenfallen. In diesen Fällen kann sich aus der epistemischen Erfüllung automatisch auch die substanzielle Erfüllung eines Desideratums ergeben, und (instrumentelles) Verstehen ermöglicht auch letzteres. Mit ‚Erfüllung von Desiderata‘ und verwandten Ausdrücken meine ich daher zumeist die epistemische Variante bzw. die Erfüllung der epistemischen Komponenten eines Desideratums. Wenn diese mit der substanziellen Erfüllung eines Desideratums zusammenfallen, ist darin allerdings auch die substanzielle Variante eingeschlossen.

Theoretische Zwecke *instrumentellen Verstehens*

Der Zweck von Verstehen in XAI-Kontexten (als Phänomen) ist also die epistemische, teils auch substanzielle Erfüllung legitimer gesellschaftlicher Desiderata, die von verschiedenen Stakeholdern an KI-Kontexte herangetragen werden. Auf dieser Grundlage können wir nun einige plausible theoretische Zwecke des Explikats formulieren: Erstens soll mein Begriff instrumentellen Verstehens besser greifbar machen, was XAI *grundsätzlich* erreichen soll. Zum einen, indem er erklärt, wieso Verstehen die oben geschilderte Mediatorfunktion zwischen Erklärbarkeit und der Erfüllung von Desiderata zukommt. Zum anderen, indem er die üblicherweise vagen Verweise auf Verstehen im Allgemeinen konkretisiert. Wenn XAI Verstehen hervorbringen soll, muss klar sein, worin dieser epistemische Zustand besteht (Krishnan, 2020; Páez, 2019; Longo et al., 2024, Kap. 3.3.3) – es braucht einen substanziellen, ‚feingetunten‘ Verstehensbegriff für XAI-Kontexte (siehe Abschnitt 4.1.1).

Zweitens soll meine Explikation einen Ansatzpunkt dafür liefern, das epistemische Ziel von XAI *in einem bestimmten Kontext* zu präzisieren. Dass die Anforderungen an Erklärungen vom konkreten Kontext abhängen, ist nicht nur eine wichtige Annahme pragmatischer Erklärungstheorien (siehe Abschnitt 3.3.1), sondern auch in der Erklärbarkeitsliteratur

⁵⁰Die epistemische Erfüllung des Desideratums ermöglicht der Endnutzerin, innezuhalten und zu überlegen, wie sie weiter vorgehen kann – auch, wenn das Desideratum substanziell unerfüllt bleibt. Plausiblerweise wechselt dann ihr Desideratum: Es geht nicht länger darum, dass das System fair ist, sondern beispielsweise darum, dass sie (trotz des unfairen Systems) eine faire Entscheidung trifft.

längst ein Allgemeinplatz (siehe etwa Arrieta et al., 2020; Langer et al., 2021; Longo et al., 2024; Miller, 2019; Páez, 2019). Offenkundig wirken sich Kontextfaktoren nicht nur auf die Adäquatheit von Verstehensmöglichkeiten, sondern auch auf das jeweils erforderliche Verstehen aus. Eine wichtige Frage ist also, was ein Stakeholder in einem bestimmten Kontext verstehen muss, um dadurch sein Desideratum erfüllen zu können. Die epistemische Erfüllung des Fairness-Desideratums unterscheidet sich *prima facie* von jener des Akkuratheits-Desideratums; eine Systementwicklerin braucht mutmaßlich ein anderes Verständnis als eine Endnutzerin oder der Gesetzgeber. Hier kommt auch die Frage danach ins Spiel, welches Verständnis von unterschiedlichen Stakeholdern legitimerweise erwartet werden kann (siehe Abschnitt 4.2.1). Mein Explikat sollte daher den Einfluss kontextueller Faktoren auf Verstehen berücksichtigen und konkretisieren, worin instrumentelles Verstehen in einem bestimmten Kontext besteht.

4.3 Adäquatheitsbedingungen des Explikats

Kommen wir nun zum zweiten Schritt des Explikations-Rezepts (siehe Abschnitt 4.1.3): der Formulierung von Adäquatheitsbedingungen, anhand derer das Explikat evaluiert werden kann. Carnap nennt vier Eigenschaften, die ein Explikat erfüllen sollte: Ähnlichkeit, Exaktheit, Fruchtbarkeit und Einfachheit (Carnap, 1971, S. 5, 7). Ich folge hingegen Georg Bruns (2016) Vorschlag, Carnaps Adäquatheitskriterien in Ähnlichkeits- und Nützlichkeitsbedingungen zu unterteilen und feingliedriger zu differenzieren. Ich diskutiere die Kriterien beider Kategorien nacheinander in den Abschnitten 4.3.1 und 4.3.2.

4.3.1 Ähnlichkeitsbedingungen

Der erste Teilschritt, Schritt (2a), besteht in der Formulierung von Ähnlichkeitsbedingungen. Die hinreichende Ähnlichkeit zwischen Explikandum und Explikat (bzw. zwischen dem alten und dem gestalteten Begriff) ist das wohl am meisten diskutierte Adäquatheitskriterium innerhalb der Begriffsgestaltungsliteratur. Der Grund dafür besteht im *Kontinuitätsproblem*: Ähneln oder gleichen sich der vorherige und der neue Begriff nicht in allen relevanten Hinsichten, so der Einwand, wechselt die Begriffsgestaltung einfach das Thema (dieser Vorwurf geht maßgeblich auf Peter Strawson, 1963, zurück). Die Begriffsgestaltung besitzt also keinen unbegrenzten ‚Gestaltungsspielraum‘: Wenn ich *instrumentelles Verstehen* als ‚Werkzeug zur Geschwindigkeitsmessung‘ definieren würde, wäre die Begriffsgestaltung gescheitert, weil keine hinreichende Kontinuität zwischen Explikandum und Explikat bestünde. Doch Kontinuität *von was* muss bestehen, damit eine Begriffsgestaltung erfolgreich ist?

Jennifer Nado unterscheidet zwischen zwei grundsätzlichen Herangehensweisen an diese Problemstellung: der semantische und der funktionale Ansatz (Nado, 2021). Dem *semantischen Ansatz* zufolge wird die erforderliche Kontinuität dadurch sichergestellt, dass das alte und neue repräsentationale Mittel sich hinsichtlich Bedeutung oder Inhalt

hinreichend ähneln (z. B. in Form extensionaler Überschneidung). Wichtig ist hierbei, dass die klassische Begriffsanalyse bei diesem Ansatz wichtige Vorarbeiten leisten kann, insofern dadurch die Bedeutung des alten Begriffs geklärt wird (vgl. ebd., S. S1513). Der *funktionale Ansatz* schlägt stattdessen vor, dass die erforderliche Ähnlichkeit durch die Kontinuität der begrifflichen Funktion sichergestellt wird (ebd., S. S1515). Nado vertritt diesbezüglich einen ‚radikalen Funktionalismus‘: Begriffe könnten diverse Funktionen, Ziele und Zwecke erfüllen, und diese könnten von den ursprünglich beabsichtigten Funktionen abweichen. Beispielsweise kommen auch eine einfache Anwendung oder Erlernbarkeit des Begriffs als relevante Ziele in Frage. Eine Begriffsgestalterin müsse sich daher fragen, ob der neue Begriff alle relevanten (nicht nur oder notwendigerweise die eigentlichen oder essenziellen) Funktionen erfüllt (ebd., S. S1521–S1522).

Ich denke, dass von Fall zu Fall entschieden werden muss, welchen Grad (und welche Form) der Ähnlichkeit eine erfolgreiche Explikation erfordert.⁵¹ In Bezug auf instrumentelles Verstehen ist die erforderliche Ähnlichkeit vornehmlich semantischer Art, weil meine Explikation einen bereits in der XAI-Literatur vorhandenen Begriff näher bestimmen soll. Autor:innen unterschiedlichster Disziplinen berufen sich auf Verstehen *simpliciter*, wenn sie das epistemische Ziel von Erklärbarkeit bzw. den Mediator zwischen XAI-Ansätzen und gesellschaftlichen Desiderata benennen (siehe Abschnitte 4.1.1 und 4.2.3). Wenn diese Autor:innen Recht haben, sollen Erklärbarkeitsansätze (und, *mutatis mutandis*, andere Verstehensmöglichkeiten) einen epistemischen Zustand hervorbringen, der vortheoretisch durch ‚Verstehen‘ und äquivalente Ausdrücke bezeichnet wird. Die Frage ist also: Was ist Verstehen *simpliciter*, und welche Aspekte dieses Alltagsbegriffs sind relevant für instrumentelles Verstehen? Ein semantisches Ähnlichkeitskriterium stellt sicher, dass es sich bei instrumentellem Verstehen tatsächlich um eine Form des Verstehens handelt. (Die funktionelle Perspektive kommt allerdings im Rahmen der Nützlichkeitskriterien ins Spiel, die ich in Abschnitt 4.3.2 aufstelle.)

Damit mein Explikat dieses Kriterium erfüllt, setze ich mich in den kommenden Kapiteln mit erkenntnis- und auch wissenschaftstheoretischen Arbeiten auseinander, die sich bemühen, die Natur von Verstehen näher zu fassen. Die Idee ist folgende: In der XAI-Literatur

⁵¹ Beispielsweise scheint semantische Ähnlichkeit nicht notwendig für begriffliche Kontinuität zu sein. Das lässt sich illustrieren, wenn man annimmt, dass semantische Ähnlichkeit in hinreichender extensionaler Überschneidung besteht. Brun nennt als Beispiel Robert Stalnakers Explikationen des Begriffs *Proposition* als (a) Funktion, die mögliche Welten in Wahrheitswerte umwandelt und (b) Menge möglicher Welten. Da Propositionen nicht zugleich Funktionen und mögliche Welten sein können, muss mindestens ein Explikat keinerlei extensionale Überlappung mit dem Explikandum aufweisen (Brun, 2016, S. 1221). Extensionale Überschneidung ist nicht einmal hinreichend für begriffliche Kontinuität. So sind *Morgenstern* und *Abendstern* trotz extensionaler Identität (zumindest einer gängigen Lesart zufolge) unterschiedliche Begriffe mit abweichender Bedeutung (Prinzing, 2018, S. 871). Mitunter arbeitet eine Begriffsgestaltung auch heraus, dass die Extension des Explikandums leer ist, und gestaltet einen neuen Begriff mit abweichender Extension (vgl. Nado, 2021, S. S1514; Prinzing, 2018, S. 872) – z. B., als Wissenschaftler:innen *Phlogiston* durch *Sauerstoff* ersetzen. Solche Fälle zeigen, dass manche Begriffsgestaltungen sogar die Absicht haben, das Thema zu wechseln (Chalmers, 2020, S. 10; Nado, 2021, S. S1514).

wird üblicherweise auf den Alltagsbegriff verwiesen, um die Mediatorrolle von Verstehen zu beschreiben (siehe Abschnitt 4.1.1). Philosoph:innen setzen sich seit einigen Jahrzehnten vermehrt mit *Verstehen* auseinander. In der Regel versuchen diese Arbeiten, mithilfe von Begriffsanalyse diesen vortheoretischen Alltagsbegriff zu definieren. Die Philosophie von Verstehen bietet daher einen guten Ausgangspunkt, um die erforderliche Ähnlichkeit zwischen Verstehen *simpliciter* und instrumentellem Verstehen zu gewährleisten.

Mein Ziel ist folglich nicht, einen Begriff zu gestalten, der radikal vom Alltagsbegriff abweicht (für einen solchen (funktionalistischen) Ansatz siehe Sleeuw, 2023). Ich möchte vielmehr eine spezielle Form des Verstehens definieren, die Verstehen *simpliciter* in für XAI-Kontexte relevanten Hinsichten semantisch ähnelt. Mein Vorgehen stützt sich daher auch auf die durch philosophische Intuitionen gestützte Analyse des Begriffs bzw. von Fallbeispielen. Die Methode der Explikation erlaubt aber ausdrücklich Abweichungen von der vortheoretischen Verwendung des Begriffs. Entsprechend muss *instrumentelles Verstehen* nicht zwangsläufig (allen) Intuitionen bezüglich der Angemessenheit von Verstehenszuschreibungen Rechnung tragen.

4.3.2 Nützlichkeitsbedingungen

Kommen wir nun zu Schritt (2b) des Explikations-Rezepts: die Formulierung von Nützlichkeitsbedingungen. Brun listet mehrere Kriterien auf, welche die theoretische Nützlichkeit des Explikats gewährleisten sollen (Brun, 2016, S. 1227–1228): Klarheit (engl. *unambiguity*), Konsistenz und Carnaps Exaktheit, Fruchtbarkeit sowie Einfachheit. Weitere potenzielle Adäquatheitskriterien sind Präzision, Messbarkeit, Geltungsbereich und Erklärungskraft. Hinzu kommen schließlich noch die theoretischen Zwecke, die das Explikat erfüllen soll.

Welche dieser Adäquatheitskriterien sind für instrumentelles Verstehen von Bedeutung? Ein relevantes Kriterium ist die Erfüllung der in Abschnitt 4.2.3 beschriebenen *theoretischen Zwecke* in den theoretischen Kontexten aus Abschnitt 4.2.2: Mein Begriff instrumentellen Verstehens soll den Alltagsbegriff von Verstehen aus der Erklärbarkeitsliteratur präzisieren und erklären, wieso Verstehen die ihm zugeschriebene Mediatorfunktion grundsätzlich erfüllen kann. Außerdem soll er in Bezug auf konkrete Kontexte erklären können, weshalb ein Desideratum durch XAI (nicht) erfüllt werden kann. Insofern ist auch *Erklärungskraft* ein relevantes Adäquatheitskriterium für instrumentelles Verstehen.

Ein weiteres Kriterium ist *Klarheit*, also die Reduzierung oder Beseitigung der Ambiguität des Explikandums durch eine möglichst eindeutige Spezifizierung der Bedeutung des Explikats. Mein Explikat soll daher die Vagheit der allgemeinen Verweise auf Verstehen innerhalb der XAI-Literatur mindern. Die verwandte *Exaktheit* verlangt, dass das Explikat mit möglichst exakten Verwendungsregeln eingeführt wird. Es sollte also so eindeutig wie möglich sein, was unter meine Definition instrumentellen Verstehens fällt. *Konsistenz* des Explikats, also Freiheit von Widersprüchen und Paradoxien, ist wohl ein Kriterium jeder Explikation.

Mein Explikat sollte außerdem *fruchtbar* sein. Carnaps Charakterisierung von Frucht-

barkeit fällt knapp aus (Dutilh Novaes, 2020, S. 1017), doch meint Fruchtbarkeit, allgemein gesprochen, dass das Explikat nützlich ist, um neue universelle Aussagen zu treffen. Carnap selbst meinte damit vor allem allgemeine empirische Gesetze oder logische Theoreme (Carnap, 1971, S. 7). Plausiblerweise kann ein Explikat aber auch deswegen nützlich sein, weil es die Formulierung universeller normativer und/oder nicht-formaler Aussagen erlaubt (vgl. Brun, 2016, S. 1224). Die Fruchtbarkeit von *instrumentellem Verstehen* ließe sich dagegen auch und gerade dadurch zeigen, dass partikuläre Aussagen über das (Nicht-)Vorhandensein von instrumentellem Verstehen in einem speziellen Kontext getroffen werden können: Um in der Praxis die Erfüllung gesellschaftlicher Desiderata zu ermöglichen und zu überwachen, muss man von Fall zu Fall entscheiden können, ob das dafür erforderliche Verstehen vom fraglichen Stakeholder erworben wurde oder nicht.

Bestenfalls wäre mein Explikat darüber hinaus auch fruchtbar für empirische und informatische Arbeiten zu Verstehen in XAI-Kontexten – etwa, indem es dazu beiträgt, bestehende Erklärbarkeitsmethoden in Hinblick auf ihre Zuträglichkeit für den Verstehenswerb empirisch zu evaluieren, oder die Entwicklung neuer Erklärbarkeitsansätze erkenntnistheoretisch zu unterfüttern. Relevant ist auch das verwandte Kriterium des großen *Geltungsbereichs*. Je mehr XAI-Kontexte sich mithilfe von instrumentellem Verstehen analysieren lassen, desto fruchtbarer ist das Explikat.

Nun, da das methodische Vorgehen geklärt ist, steige ich in den folgenden Kapiteln in die philosophische Literatur zum Thema Verstehen *simpliciter* ein. Den Anfang macht die Debatte zur Gradierbarkeit von Verstehen.

5 Verstehensgrade

Verstehen ist gradierbar: Ich verstehe die LIME-Methode besser als meine Mutter, aber schlechter als die LIME-Entwickler Marco Ribeiro, Sameer Singh und Carlos Guestrin. Und zu Beginn meiner Promotion verstand ich viel schlechter als heute, warum sich tiefe neuronale Netze von einer Sackgasse der KI-Forschung zum technologischen (Fast-)Alleskönner gemausert haben. Dennoch ist mein Verständnis der historisch-technischen Zusammenhänge keinesfalls perfekt. Natürlich gibt es auch absolute Zuschreibungen von Verstehen – etwa, wenn jemand sagt, dass er jetzt endlich die LIME-Methode versteht. Doch selbst in solchen Fällen scheint unbestreitbar, dass das jeweilige Verständnis schlechter oder besser sein könnte, als es ist (Malfatti, 2023, Kap. 1.4.3). Tatsächlich ist Gradierbarkeit eine der wenigen Eigenschaften von Verstehen, die innerhalb der philosophischen Forschung zum Thema als konsensuell gelten können.⁵²

Die Gradierbarkeit von Verstehen wirft mindestens zwei Fragen auf. Die erste ist die Frage danach, wie man Verstehensgrade in einer Definition bestmöglich abbilden kann. Im Anschluss an Christoph Baumberger lassen sich diesbezüglich drei Strategien unterscheiden (Baumberger, 2019, Kap. 2.3; vgl. auch Kelp, 2015, S. 3805): Die erste Strategie nimmt minimales Verstehen als Ausgangspunkt. Verstehensgrade können dann als Verbesserungen dieses minimalen Verstehens definiert werden. Die zweite Strategie bestimmt stattdessen maximales Verstehen und beschreibt Verstehensgrade relativ zu diesem Idealzustand. Die dritte Strategie besteht darin, unmittelbar zu bestimmen, worin ein bestimmter Verstehensgrad bzw. besseres Verstehen besteht.⁵³

Relevant für alle drei Strategien ist die zweite Frage danach, *in welcher Hinsicht* der Grad des Verstehens variieren kann. Welche gradierbaren Aspekte von Verstehen determinieren den Verstehensgrad, den eine Person besitzt? Um diese zweite Frage soll es im vorliegenden Kapitel gehen. Ich argumentiere, dass instrumentelles Verstehen – anders als

⁵²Dieser Konsens betrifft streng genommen lediglich interrogatives Verstehen (Verstehen-warum, Verstehen-wie, Verstehen-wo...) und objektuales Verstehen (z. B. das Verstehen eines Themengebiets). Propositionales Verstehen, also Verstehen-dass, ist der Standardauffassung zufolge nicht gradierbar. In Kapitel 6 schlage ich allerdings vor, dass auch propositionales Verstehen als gradierbar aufgefasst werden sollte.

⁵³Definitionen minimalen Verstehens bieten beispielsweise Christoph Baumberger (2019, S. 376) und Kareem Khalifa (2017, S. 14). Maximales oder ideales Verstehen betrachten z. B. Mario Hubert & Federica Malfatti (2023, S. 590); Christoph Kelp (2015, S. 3811) und abermals Kareem Khalifa (2017, S. 4). Die dritte Strategie, Verstehensgrade unmittelbar zu bestimmen, findet sich bei Christoph Baumberger (2019, S. 376); Miguel Egler (2021, S. 6069); Kareem Khalifa (2017, S. 14); Amber Riaz (2015, S. 122) oder Daniel Wilkenfeld (2017a, S. 1275–1276). Einige Vor- und Nachteile dieser drei Herangehensweisen diskutiert Baumberger (2019, S. 372).

Verstehen *simpliciter* – nicht gradierbar ist. Allerdings besteht instrumentelles Verstehen in einem kontextspezifischen Verstehensgrad. Um diesen Verstehensgrad bestimmen zu können, muss klar sein, wodurch sich Verstehensgrade voneinander unterscheiden. Damit ähnelt mein Vorschlag der dritten Herangehensweise an die erste Frage, insofern ich unmittelbar bestimme, worin ein bestimmter Verstehensgrad besteht bzw. was besseres oder schlechteres Verstehen ausmacht.

Diese Stoßrichtung bestimmt den Aufbau des restlichen Kapitels: In Abschnitt 5.1 unterscheide ich vier Dimensionen, entlang derer der Grad des Verstehens variieren kann. Anschließend konkludiere ich in Abschnitt 5.2, dass instrumentelles Verstehen nicht gradierbar ist und dass die Unterscheidung von Verstehensgraden trotzdem wichtig ist, um instrumentelles Verstehen zu charakterisieren.

5.1 Dimensionen der Gradierbarkeit von Verstehen

In diesem Abschnitt möchte ich näher fassen, entlang welcher Dimensionen der Grad des Verstehens variieren kann. Auf Grundlage bestehender Vorschläge unterscheide ich vier übergeordnete Dimensionen: Vollständigkeit (Abschnitt 5.1.1), Faktizität (Abschnitt 5.1.2), Begreifen (Abschnitt 5.1.3) sowie Rechtfertigung (Abschnitt 5.1.4). Das heißt: Je vollständiger oder faktiver das Überzeugungssystem einer verstehenden Person ist, je besser diese Person relevante Zusammenhänge begreift oder je besser ihre Überzeugungen gerechtfertigt sind, desto höher ist, *ceteris paribus*, auch ihr Verstehensgrad. Eine Übersicht bietet Abbildung 5.1. Da sich die Dimensionen der Gradierbarkeit nicht immer klar auseinanderhalten lassen, gehe ich abschließend noch auf Überschneidungen zwischen den einzelnen Dimensionen ein (Abschnitt 5.1.5).

5.1.1 Dimension 1: Vollständigkeit

Die erste Dimension der Gradierbarkeit von Verstehen betrifft den Grad, zu dem sich eine Person vollständigem Verstehen annähert. Gemeint ist hier inhaltliche Vollständigkeit: Nehmen wir an, die Inhalte eines Themengebiets G (z. B. die menschliche Evolution) bestehen in den Propositionen p_1 , p_2 und p_3 . Um G vollständig zu verstehen, braucht eine Person entsprechend die Überzeugung, dass p_1 , dass p_2 und dass p_3 . Im Gegensatz dazu glaubt eine Person mit unvollständigem Verstehen beispielsweise nur, dass p_1 und p_3 , oder p_2 und p_3 . Dies ermöglicht eine unkomplizierte Bestimmung verschiedener Grade der Vollständigkeit: Man muss bloß den Anteil der Propositionen angeben, die zu G gehören und vom Verstehenssubjekt geglaubt bzw. nicht geglaubt werden.

Diese einfache Zählung führt natürlich dazu, dass eine Person A , die fünf nebensächliche Propositionen über G glaubt, G ebenso vollständig versteht wie Person B , die fünf Propositionen glaubt, die sehr wichtig für das Themengebiet sind. Das stimmt nicht zwangsläufig mit unserer intuitiven Einschätzung überein. Nehmen wir das Verstehen der menschlichen Evolution: Möglicherweise wirkt sich die Überzeugung, dass der moderne Mensch aus

Arten der Gattung *Australopithecus* hervorging, stärker auf die Vollständigkeit meines Verstehens dieses Themengebiets aus als die Überzeugung, dass der erste *Homo erectus* 81.237 Kopfhare besaß.⁵⁴ Möchte man eine gewichtete Vollständigkeit von Verstehen bestimmen, ist eine Unterscheidung zwischen zentralen und peripheren Überzeugungen hilfreich. Ich komme darauf in Abschnitt 8.3.3 zurück, wenn es um die Faktizität von Verstehen geht.

Die Dimension der Vollständigkeit diskutiert auch Christoph Kelp im Rahmen seiner Auseinandersetzung mit Verstehensgraden. Er unterscheidet diese anhand des Grads, zu dem sich Verstehen vollumfänglichem bzw. vollständigem (engl. *fully comprehensive*), maximal gut verknüpftem Wissen bezüglich eines Phänomens annähert (Kelp, 2015, S. 3812; eine jüngere Darstellung bieten Kelp, 2017 und 2021, Kap. 4). Eine Person besitzt vollständiges Wissen über ein Phänomen, wenn sie alles weiß, was es darüber zu wissen gibt (Kelp, 2015, S. 3809–3810). Vollständigkeit ist aber laut Kelp nicht hinreichend: Zusätzlich muss die maximal verstehende Person bestimmte Propositionen auch *aufgrund* der (z. B. logischen, explanatorischen oder probabilistischen) Relationen zwischen den Propositionen glauben – etwa, wenn einzelne Propositionen eine andere implizieren (ebd., S. 3810). Diese Verknüpftheit bildet entsprechend eine zweite Dimension der Gradierbarkeit.

Ich greife die Dimension der Verknüpftheit in Abschnitt 5.1.3 erneut auf. Betrachten wir zunächst die Vollständigkeitsdimension genauer. Denn die Propositionen, die ein Themengebiet oder ein anderes Objekt des Verstehens ausmachen, können weiter ausdifferenziert werden. Ich stütze mich dafür auf die Dimensionen der Gradierbarkeit, die Catherine Elgin unterscheidet: Breite, Tiefe, Relevanz und Faktizität (Elgin, 2007, S. 36–37). Auf den Grad der Faktizität komme ich in Abschnitt 5.1.2 zurück. An dieser Stelle möchte ich vorschlagen, dass sich Elgins übrige Dimensionen der Breite, Tiefe und Relevanz als Subdimensionen der Vollständigkeit auffassen lassen.⁵⁵ Wenn eine Person *A* beispielsweise mehr Details von Themengebiet *G* erfasst als eine andere Person *B* (Tiefe); oder wenn *A* einen besseren Überblick über *G* hat als *B* (Breite); oder wenn *A* die Relevanz einzelner Aspekte von *G* akkurater einschätzt als *B* (Relevanz); dann besitzt *A ceteris paribus* einen höheren Verstehensgrad als *B*, weil *A*s Verstehen vollständiger ist.

Das lässt sich am Beispiel des Verstehens der menschlichen Evolution illustrieren. Manche der zugehörigen Propositionen betreffen eher übergeordnete Gesichtspunkte – z. B. wenn jemand viele unterschiedliche Aspekte der menschlichen Evolution grob versteht oder die menschliche Evolution in die Entwicklung von Lebewesen im Allgemeinen einordnen kann. Eine solche Person besitzt ein *breites* Verstehen der menschlichen Evolution. Andere Propositionen beschreiben eher spezifische Details der menschlichen Entwicklung. Wenn jemand beispielsweise die Evolution des menschlichen Auges sehr detailliert versteht, dann besitzt diese Person ein *tiefes* Verständnis. Manche der Propositionen, die

⁵⁴Dies berührt die Problematik trivialer Wahrheiten bzw. das Clutter-Avoidance-Prinzip (Harman, 1986). Siehe etwa Elgin, 2017, S. 10–11; Khalifa, 2017, Kap. 4 und Søndergaard, 2023, S. 2619 für Bemerkungen dazu im Kontext von Verstehen.

⁵⁵Damit weiche ich von Kelps Position ab, siehe Fußnote 60.

das Themengebiet betreffen, sind hingegen ‚Meta-Propositionen‘ über die *Relevanz* der einzelnen Aspekte der menschlichen Evolution. Jemand, der hinsichtlich dieser Subdimension einen hohen Verstehensgrad besitzt, kann z. B. die Relevanz klimatischer Faktoren in Relation zur Bedeutung spontaner Genmutationen für die Entwicklung des Menschen korrekt einschätzen.

Diese drei Subdimensionen bilden intuitiv plausible Unterschiede zwischen Verstehensgraden in XAI-Kontexten ab. Betrachten wir dazu exemplarisch Ansätze, die auf Bildklassifikationssysteme angewandt werden (Abschnitt 3.2.2): Lokale Attributionsmethoden geben Aufschluss über eine einzelne Klassifikation und erhellen dadurch ein spezifisches Detail des Bildklassifikationssystems. Das resultierende Verstehen ist daher eher *tief* als z. B. breit. Bei globalen Ansätzen kann es dagegen genau umgekehrt sein. Beispielsweise bietet die Feature-Visualisierung einer Bildklasse eine Übersicht über typische Merkmale von Mitgliedern dieser Klasse, ohne auf einzelne Klassifikationen einzugehen. Das resultierende Verstehen ist dadurch nicht besonders tief, aber dafür *breiter* als bei lokalen Methoden. Schließlich können Salienzkarten den Einfluss verschiedener Pixel auf die Klassifikation durch unterschiedliche Farbtöne kennzeichnen (siehe etwa Abbildung 3.2b). Das ermöglicht es, zu verstehen, wie *relevant* die einzelnen Bildmerkmale relativ zueinander für die fragliche Klassifikation waren.

Zusammengefasst: Verstehensgrade können in Abhängigkeit von der inhaltlichen Vollständigkeit der geglaubten Propositionen variieren. Innerhalb dieser übergeordneten Dimension lassen sich mit Breite, Tiefe und Relevanz drei Subdimensionen der Vollständigkeit unterscheiden – je nachdem, ob die geglaubten Propositionen übergeordnete Aspekte, Details oder die Relevanz anderer Propositionen des verstandenen Themengebiets (o. ä.) betreffen.

5.1.2 Dimension 2: Faktizität

Die zweite Dimension der Gradierbarkeit betrifft die Faktizität der in Verstehen involvierten Überzeugungen. Die Idee ist hier, dass eine Person umso besser versteht, desto mehr ihr Überzeugungssystem der Wahrheit entspricht. Eine Person, die ein maximal faktives Verstehen der menschlichen Evolution besitzt, hat demnach ausschließlich und vollkommen wahre Überzeugungen über dieses Themengebiet.

Auch hier möchte ich mehrere Subdimensionen unterscheiden. Die erste Subdimension hängt mit dem Anteil an wahren bzw. falschen Überzeugungen zusammen: Wenn *A* mehr falsche Überzeugungen über Themengebiet *G* besitzt als *B*, dann hat *A* *ceteris paribus* einen geringeren Verstehensgrad hinsichtlich *G* als *B*. Betrachten wir exemplarisch eine Proposition (p_1) des Themengebiets der menschlichen Evolution: *Arten der Gattung Australopithecus bilden die Vorfahren der Gattung Homo*. Wir können der Einfachheit halber erneut annehmen, dass dieses Themengebiet aus der Menge an Propositionen $\{p_1, p_2, p_3\}$ besteht. Jemand, der beispielsweise glaubt, dass $\neg p_1$ und dass p_2 und dass p_3 , besitzt dann einen geringeren Verstehensgrad als jemand, der p_1 – p_3 für wahr hält – schließlich besitzt

eine solche Person eine falsche Überzeugung.

Wie auch bei der Vollständigkeitsdimension könnte es sein, dass es zentrale und periphere Überzeugungen gibt, die sich jeweils stärker bzw. schwächer auf den Faktizitätsgrad auswirken. Die oben genannte Proposition p_1 ist im Vergleich zu einer anderen Proposition p_2 (z. B. *Das Grundprinzip der menschlichen Evolution ist die natürliche Auslese aufgrund des Überlebensvorteils von Individuen mit an die Umwelt angepassten Genvarianten*) vielleicht eher peripher. Wenn jemand p_2 für falsch hält, reduziert dies den Verstehensgrad womöglich stärker als wenn sich jemand in Bezug auf p_1 irrt. Ich vertage diese Frage auf Abschnitt 8.3.3.

Die zweite Subdimension der Faktizität von Verstehen übernehme ich abermals von Catherine Elgin. In Abschnitt 5.1.1 habe ich erwähnt, dass sie neben Breite, Tiefe und Relevanz auch noch die Dimension der Faktizität unterscheidet (Elgin, 2007, S. 37). Diese Dimension besteht in dem Grad, zu dem sich einzelne Überzeugungen der Wahrheit annähern: Wenn A s Überzeugungen über G die Wahrheit besser approximieren als jene von B , dann besitzt A *ceteris paribus* einen höheren Verstehensgrad hinsichtlich G als B . Wenn A beispielsweise glaubt, dass sich der *Homo erectus* vor exakt 2 Millionen Jahren entwickelte, obwohl dies in Wahrheit vor 2,04 Millionen Jahren geschah, dann ist A s Verstehensgrad *ceteris paribus* geringer als von Person B , die glaubt, dass der *Homo erectus* vor exakt 2,03 Millionen Jahren entstand.⁵⁶ Ein anderes Beispiel ist der wissenschaftliche Fortschritt – etwa, wenn im Rahmen wissenschaftlicher Forschung die Evolution allmählich besser verstanden wird.⁵⁷

Die beiden Subdimensionen der Faktizität lassen sich auch auf das Verstehen in XAI-Kontexten übertragen. Beginnen wir mit der ersten Subdimension, die mit dem Anteil an falschen Überzeugungen zu tun hat. Wir können uns ein Worst-Case-Szenario vorstellen, in dem eine Post-hoc-Methode vollkommen unabhängig von dem tatsächlichen Verhalten des zu erklärenden KI-Systems ist und dessen Klassifikationskriterien völlig falsch darstellt (siehe Abschnitt 3.2.1). In einem solchen Szenario könnte jemand z. B. zu der falschen Überzeugung gelangen, dass Merkmal x der einzige Grund für eine bestimmte Klassifikation war, obwohl x in Wahrheit keinen Einfluss darauf hatte. Diese falsche Überzeugung würde den Verstehensgrad entlang der Faktizitätsdimension reduzieren – womöglich so weit, dass wir gar nicht mehr von Verstehen sprechen würden.

In anderen Fällen sind die Überzeugungen, die jemand auf Basis einer XAI-Erklärung erwirbt, nicht völlig korrekt, sondern nur ungefähr wahr. Hier kommt die zweite Sub-

⁵⁶Das Beispiel ist angelehnt an einen ähnlichen Fall von Eva Schmidt (2026, S. 229). Es ist nicht trivial, diese Approximation an die Wahrheit (engl. *verisimilitude*) theoretisch zu fassen. Für eine Diskussion verschiedener Ansätze siehe Oddie & Cevolani (2022). Ich beschränke mich hier auf das intuitive Verständnis approximativer Wahrheit, das durch Beispiele wie das im Text genannte illustriert wird.

⁵⁷Wir können die Dimensionen der Vollständigkeit und Faktizität auseinander halten, indem wir zulassen, dass auch partielles Verstehen maximal faktiv sein kann: Wenn alle Überzeugungen einer Person über Themengebiet G (o. ä.) vollkommen wahr sind, ist ihr Verständnis maximal faktiv – auch dann, wenn ihre Überzeugungen nur eine kleine Teilmenge der Propositionen von G abdecken und ihr Verstehen daher unvollständig ist.

dimension der Faktizität ins Spiel. In Abschnitt 2.3.1 habe ich aufgezeigt, dass manche KI-Systeme epistemisch opak sind, weil sie auf fremdartigen Repräsentationen beruhen. Gleichzeitig wurde dort deutlich, dass mitunter eine näherungsweise Zuordnung dieser Repräsentationen zu menschenverständlichen Begriffen gelingt. Beispielsweise könnte der Begriff *Schlappohr* in bestimmten Hinsichten einer Gruppe von Merkmalen entsprechen, die ein Bildklassifikationssystem verwendet, um Hunde zu erkennen (vgl. Olah et al., 2018) – auch wenn das System den Begriff eigentlich nicht besitzt und z. B. manchmal auch Pixelgruppen dazu zählt, die Menschen als ‚Lefze‘ bezeichnen würden, oder die aus menschlicher Perspektive keinen Sinn ergeben. Jemand, der glaubt, dass das System Hunde an ihren Schlappohren erkennt, hat in einem solchen Fall eine Überzeugung, die nicht ganz, aber ungefähr wahr ist.

Zusammengefasst: Verstehensgrade können in Abhängigkeit von der Faktizität der Überzeugungen über das verstandene Themengebiet (o. ä.) variieren. Innerhalb dieser übergeordneten Dimension lassen sich zwei Subdimensionen unterscheiden: zum einen der Anteil an falschen Überzeugungen, zum anderen der Grad, zu dem eine einzelne Überzeugung die Wahrheit approximiert.

5.1.3 Dimension 3: Begreifen

Die nächste Dimension übernehme ich von Christoph Baumberger. Er schlägt vor, objektuales Verstehen eines Themengebiets mithilfe einer Theorie anhand von vier Bedingungen zu explizieren (Baumberger, 2019, S. 376): Die erste Bedingung besagt, dass sich das verstehende Subjekt hinreichend auf die fragliche Theorie (z. B. die Evolutionstheorie) verpflichten muss (engl. *commitment*). Diese Verpflichtung ist zwar gradierbar, beeinflusst aber den Verstehensgrad nicht. Die drei weiteren Bedingungen sind dagegen evaluativ: Eine davon betrifft den Grad der Richtigkeit der Theorie, die Baumberger durch deren relevante Ähnlichkeit zum Theoriegegenstand ausbuchstabiert. Er betrachtet also, anders als ich (siehe Abschnitt 5.1.2), nicht die Faktizität der Verstehensüberzeugungen, sondern jene der verstandenen Theorie. Ich werde in Abschnitt 8.3 und Kapitel 12 darauf zurückkommen, warum ich diese Herangehensweise für ungeeignet halte, um die Faktizität von Verstehen zu charakterisieren. Die nächste evaluative Bedingung beschreibt den Grad, zu dem die Verpflichtung des verstehenden Subjekts auf die Theorie gerechtfertigt ist. Dieser Dimension widme ich mich in Abschnitt 5.1.4. Die hier relevante evaluative Bedingung spezifiziert den Grad, zu dem das verstehende Subjekt die Theorie begreift. Dieses Begreifen von Zusammenhängen ist die dritte Dimension der Gradierbarkeit von Verstehen, die ich unterscheiden möchte.

Diesbezüglich lassen sich abermals zwei Subdimensionen ausmachen. Die erste klang bereits in Christoph Kelps Dimension der Verknüpftheit aus Abschnitt 5.1.1 an. Zur Erinnerung: Damit das Wissen einer verstehenden Person gut verknüpft ist, so Kelp, muss sie die gewussten Propositionen auch aufgrund der dazwischen bestehenden Relationen glauben (Kelp, 2015, S. 3810). Beispielsweise weiß eine Person, welche die menschliche Evolution

verstehen, nicht nur, dass (a) sich die frühen Vorfahren des modernen Menschen aufrecht auf Bäumen entlanghangelten, dass (b) eine Trockenperiode zum Rückgang afrikanischer Regenwälder führte und dass (c) die Vorfahren des heutigen Menschen den aufrechten Gang auf dem Boden perfektionierten. Ihre Überzeugung, dass (c), basiert zudem auf (a) und (b). Kelps Anforderung trägt einer verbreiteten Intuition Rechnung: Um zu verstehen, genügt es nicht, voneinander isolierte Fakten zu glauben. Die geglaubten Propositionen müssen auch angemessen miteinander verknüpft sein. Anders ausgedrückt: Wenn *A* Überzeugungen über ein Themengebiet *G* (o. ä.) besser durch z. B. inferenzielle Relationen miteinander verknüpft und dadurch weniger isoliert voneinander sind als jene von *B*, dann besitzt *A ceteris paribus* einen höheren Verstehensgrad als *B*, weil *A* besser begreift, wie die einzelnen Aspekte von *G* miteinander zusammenhängen. Mehr dazu folgt in Kapitel 6.

Die zweite Subdimension stammt wiederum von Baumberger. Er konkretisiert das für Verstehen charakteristische Begreifen in Form verstehensbezogener Fähigkeiten (Baumberger, 2019, Kap. 4.2): Wer die menschliche Evolution zu einem hohen Grad versteht, kann die Evolutionstheorie z. B. auf andere Spezies als die Hominiden anwenden oder mehr Schlüsse hinsichtlich der Auswirkungen kontrafaktischer Lebensbedingungen der Vorfahren des *Homo sapiens* auf die menschliche Evolution ziehen. Die Idee ist also folgende: Wenn *A* über mehr oder ausgeprägtere verstehensbezogene Fähigkeiten hinsichtlich *G* verfügt als *B* (z. B., weil *A* mehr Aspekte von *G* erklären kann), dann besitzt *A ceteris paribus* einen höheren Verstehensgrad als *B*, weil *A* besser begreift, wie die einzelnen Aspekte von *G* miteinander zusammenhängen. Ich betrachte diesen Aspekt von Verstehen in Kapitel 9.

Übertragen wir die beiden Subdimensionen des Begreifens auf das Verstehen in XAI-Kontexten. Beispielsweise kann die Kombination mehrerer Erklärbarkeitsmethoden (Olah et al., 2018) Verknüpfungen zwischen vormals isolierten Überzeugungen besonders gut aufzeigen. Womöglich glaubt man zunächst (isoliert voneinander), dass ein Bildklassifikationssystem ein bestimmtes Bild als ‚Hund‘ klassifiziert hat – z. B. aufgrund einer Salienzkarte – und dass das System Hunde an hängenden Ohren sowie Katzen an spitzen Ohren erkennt – z. B. auf Grundlage einer globalen Methode (ebd.). Durch die Kombination mehrerer XAI-Ansätze (z. B. Feature-Visualisierung und Attributionsmethoden) könnte man begreifen, wie sich die spezifische Form des Ohrs auf die Klassifikation als ‚Hund‘ bzw. ‚Katze‘ auswirkt. Dadurch könnte man beispielsweise glauben, dass ein bestimmtes Bild als ‚Hund‘ klassifiziert wurde, *auf der Basis*, dass das Klassifikationssystem Schlappohren, aber keine spitzen Ohren detektiert hat.

Andere Methoden eignen sich besonders gut für den Erwerb bestimmter verstehensbezogener Fähigkeiten. Beispielsweise können kontrafaktische Erklärungen (Abschnitt 3.2.2) die Fähigkeit eines Verstehenssubjekts verbessern, kontrafaktische Schlüsse darüber zu ziehen, welche Änderungen der Eingabedaten zu welchen Veränderungen in der Ausgabe führen würden – beispielsweise darüber, was passiert, wenn ein Bild einen Hund mit spitzen Ohren zeigt. Das Ziehen solcher Schlussfolgerungen ist eine der Fähigkeiten, die charakteristisch für Verstehen ist (mehr dazu in Abschnitt 9.2.5).

Zusammengefasst: Verstehensgrade können in Abhängigkeit davon variieren, wie gut die

verstehende Person Zusammenhänge zwischen Aspekten des verstandenen Themengebiets (o. ä.) begreift. Dieses Begreifen lässt sich anhand zweier Subdimensionen näher bestimmen: einerseits die Verknüpftheit der Überzeugungen untereinander, andererseits das Verfügen über verstehensbezogene Fähigkeiten.

5.1.4 Dimension 4: Rechtfertigung

Die letzte Dimension der Gradierbarkeit von Verstehen, die ich diskutieren möchte, betrifft den Grad der Rechtfertigung der in Verstehen involvierten Überzeugungen. Die Grundidee lautet, dass eine Person *ceteris paribus* ein Themengebiet *G* (o. ä.) umso besser versteht, desto besser ihre Überzeugungen über *G* gerechtfertigt sind. Vergleichen wir abermals zwei Personen *A* und *B*. Person *A* stützt ihre (wahre) Überzeugung, dass der moderne Mensch aus Arten der Gattung *Australopithecus* hervorging, auf das Zeugnis einer renommierten Professorin der Evolutionsbiologie. Person *B* besitzt dieselbe Überzeugung wie *A*, doch stützt sie sich dafür auf ein dubioses Youtube-Video. Intuitiv ist das Verständnis von *A* besser als jenes von *B*, weil *A*s Überzeugung besser gerechtfertigt ist (diese Intuition scheint auch z. B. Kelp, 2017, zu teilen).

Lassen sich auch Subdimensionen der Rechtfertigung von Verstehensüberzeugungen unterscheiden? Die Antwort auf diese Frage hängt von der bevorzugten Theorie der Rechtfertigung ab (mehr dazu in Kapitel 10). Auf Grundlage der theoretischen Tugenden von Thomas Kuhn (siehe Abschnitt 3.2.4) unterscheidet Baumberger fünf Subdimensionen von Rechtfertigung, die den Verstehensgrad beeinflussen (Baumberger, 2019, S. 382–383): Die erste Subdimension (a) betrifft die interne Konsistenz bzw. den Grad der externen Kohärenz der Theorie mit Hintergrundannahmen. Subdimension (b) bestimmt den Grad der Übereinstimmung der Theorie mit verfügbaren Evidenzen. Subdimension (c) entspricht dem Grad, zu dem die Theorie der Erreichung zusätzlicher epistemischer Ziele (z. B. Erklärungskraft, Fruchtbarkeit) zuträglich ist. Subdimension (d) ist internalistisch und besteht in dem Grad, zu dem das Subjekt beurteilen kann, wie gut (a)–(c) erfüllt sind. Schließlich ist Subdimension (e) externalistisch, insofern sie den Grad der Reliabilität des Prozesses charakterisiert, den das epistemische Subjekt für die Evaluation der Theorie gemäß Subdimension (d) verwendet.⁵⁸

Baumbergers Vorschlag lässt sich mit kleineren Anpassungen auch auf das Verstehen in XAI-Kontexten übertragen. Eine unmittelbare Anwendung ist nicht möglich, weil sich

⁵⁸Mehr zu externalistischer bzw. internalistischer Rechtfertigung folgt in Abschnitt 10.2.1. Baumberger behält sich vor, dass manche Subdimensionen den Verstehensgrad nicht beeinflussen: Es sei uneindeutig, ob sich der Grad der Stützung einer Theorie durch die vorhandenen Daten auf den Verstehensgrad auswirke (Baumberger, 2019, S. 384) – das betrifft mutmaßlich Subdimension (b), also die Übereinstimmung der Theorie mit verfügbaren Evidenzen. Ich bin unsicher, weshalb er diese Einschränkung vornimmt. Wenn man eine evidenziell schwach gestützte Theorie für wahr hält, ist die resultierende Überzeugung (*ceteris paribus*) nur schwach gerechtfertigt. Folgt man Baumberger darin, dass der Verstehensgrad durch den Rechtfertigungsgrad beeinflusst wird, scheint es sich bei dem genannten Szenario um einen Parade Fall schwacher Rechtfertigung zu handeln, die auch den Verstehensgrad mindern müsste.

Baumbergers Subdimensionen auf das Verstehen eines natürlichen Phänomens mithilfe einer Theorie beziehen. XAI-Erklärungen spielen in Erklärbarkeitskontexten allerdings eine Rolle analog zu jener von Theorien in dem Sinne, dass sie ebenfalls als Vehikel für das Verstehen eines ‚Phänomens‘ (z. B. die Ausgabe eines KI-Systems) dienen. Demzufolge wären die Überzeugungen eines verstehenden Subjekts besser gerechtfertigt, wenn (a) die XAI-Erklärung in sich konsistent ist und je stärker sie mit zusätzlichen Hintergrundannahmen kohäriert, (b) je besser die Erklärung mit sonstigen Evidenzen (z. B. anderen Erklärungen) kompatibel ist, (c) je besser die Erklärung zu den epistemischen Zielen des verstehenden Subjekts passt (z. B. Einfachheit), (d) je besser das verstehende Subjekt fähig ist, (a)–(c) zu beurteilen und (e) je zuverlässiger dieser Evaluationsprozess ist.

Ich werde in Abschnitt 10.2.3 einen alternativen Vorschlag zur Rechtfertigung von Verstehensüberzeugungen machen. Demzufolge sollte Rechtfertigung im Kontext von Verstehen internalistisch aufgefasst und am besten durch einen hybriden Ansatz charakterisiert werden, der kohärentistische und fundamentalistische Elemente miteinander kombiniert. Daraus ergeben sich zwei Subdimensionen von Rechtfertigung: zum einen die (fundamentalistische) Evidenz des verstehenden Subjekts, zum anderen die Kohärenz seiner Überzeugungen (vgl. Schmidt, 2026).

Auch diese beiden Subdimensionen können auf XAI-Kontexte übertragen werden. Die Rechtfertigung von Verstehensüberzeugungen und damit auch der Verstehensgrad wird demnach zum einen durch die verfügbaren Evidenzen beeinflusst – beispielsweise dadurch, ob jemand seine Überzeugungen auf eine zuverlässige Erklärbarkeitsmethode oder vielmehr das Zeugnis eines Scharlatans stützt. Zum anderen kommt es auch darauf an, wie stark die Verstehensüberzeugungen miteinander kohärieren und einander stützen. Wenn jemand glaubt, dass ein Katzenbild aufgrund der spitzen Ohren als ‚Katze‘ klassifiziert wurde und auch, dass das Tier auf dem fraglichen Foto Schlappohren aufweist, reduziert oder zerstört dies den Grad der Rechtfertigung, weil diese beiden Überzeugungen inkohärent sind.

Zusammengefasst: Verstehensgrade können in Abhängigkeit vom Grad der Rechtfertigung der darin involvierten Überzeugungen variieren. Welche Subdimensionen Rechtfertigung aufweist, hängt von der jeweils favorisierten Theorie der Rechtfertigung ab. Eine Möglichkeit besteht darin, eine fundamentalistisch-evidenzialistische und eine kohärentistische Subdimension zu unterscheiden.

5.1.5 Überschneidungen zwischen den einzelnen Dimensionen

In den vorherigen Abschnitten habe ich vier Dimensionen unterschieden, entlang derer der Verstehensgrad variieren kann. Jede dieser Dimensionen weist mehrere Subdimensionen auf. Insgesamt ergibt sich daraus die in Abbildung 5.1 dargestellte Taxonomie von Dimensionen der Gradierbarkeit von Verstehen. Die vorgeschlagenen Unterscheidungen sind allerdings nicht immer trennscharf.

Zum einen lassen sich die Subdimensionen der Vollständigkeit nicht immer klar voneinander unterscheiden. Beispielsweise gibt es nicht immer eine klare Grenze zwischen

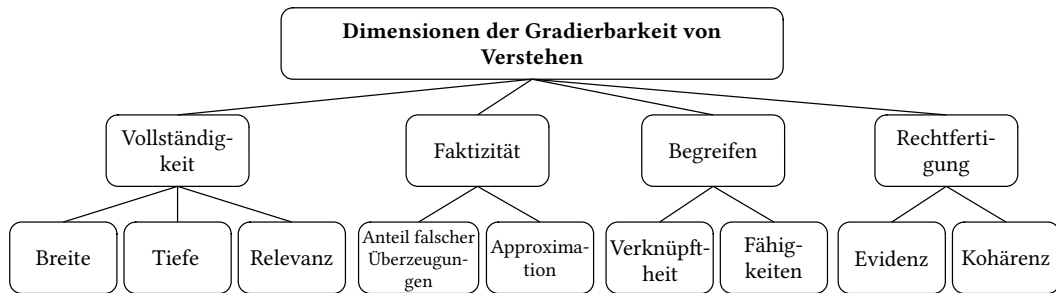


Abbildung 5.1: Die von mir vorgeschlagene Systematisierung der Dimensionen der Gradierbarkeit von Verstehen – Vollständigkeit, Faktizität, Begreifen und Rechtfertigung – samt zugehöriger Subdimensionen. Bildquelle: Eigene Darstellung. Siehe Teil IV für Details.

übergeordneten Aspekten eines Themengebiets (o. ä.) und einzelnen seiner Details, und die Relevanz einzelner Aspekte zu kennen gehört auch zu einem tiefen Verständnis dazu. Nehmen wir das Bildklassifikationssystem, das Hunde an Schlappohren erkennt. Handelt es sich dabei um einen Aspekt, der einem tiefen Verständnis des Systems entspricht? Immerhin geht es um die Klassifikation einer einzelnen von womöglich hunderten Bildklassen. Im Vergleich zur übergeordneten Funktionsweise oder Architektur des Systems ist das Schlappohr-Kriterium ein spezifisches Detail. Andererseits stellt solch ein globales Klassifikationskriterium eine Generalisierung dar, die das lokale Klassifikationsverhalten des Systems vereinfacht. Man könnte gleichermaßen die Position vertreten, dass die Proposition, dass das System Hundebilder an Schlappohren erkennt, zu einem eher breiten Verständnis des Systems gehört. Und wenn man versteht, dass Schlappohren relevanter für die Detektion von Hunden sind als Schnauzen, vertieft und verbreitert man damit nicht auch zugleich sein Verständnis? Breite, Tiefe und Relevanz gehen fließend ineinander über und es ist oft schwierig, eine Proposition eindeutig einer dieser Subdimensionen zuzuordnen.

Zum anderen spielt die Dimension des Begreifens von Zusammenhängen in diverse andere Dimensionen der Gradierbarkeit hinein. Erstens hängt die Verknüpftheit von Verstehensüberzeugungen mit der Faktizität und Vollständigkeit von Verstehen zusammen. Denn damit Verstehen faktiv und vollständig ist, müssen die kognitiv begriffenen Zusammenhänge zwischen geglaubten Propositionen auch die tatsächlichen Relationen widerspiegeln. Zweitens ist die Verknüpfung von Überzeugungen bzw. begriffenen Propositionen konstitutiv für Kohärenz und damit für eine Subdimension von Rechtfertigung. Je kohärenter und damit besser gerechtfertigt die Überzeugungen einer Person sind, desto besser sind diese Überzeugungen miteinander verknüpft und – sofern diese Verknüpfungen akkurat sind – desto besser begreift diese Person die Zusammenhänge zwischen einzelnen Aspekten des verstandenen Themengebiets (o. ä.). Und drittens finde ich es plausibel anzunehmen, dass verstehensbezogene Fähigkeiten aus einem adäquat verknüpften Netzwerk an Überzeugungen erwachsen. Am deutlichsten wird dies bei inferenziellen Fähigkeiten: Um aus

Prämissen eine Konklusion abzuleiten, muss ich diese Prämissen und die Konklusion glauben. Um etwas erklären zu können, muss ich glauben, dass p weil q . Wenn dem so ist, korrelieren die beiden Subdimensionen von Begreifen miteinander.

5.2 Instrumentelles Verstehen ist nicht gradierbar

Entgegen des breiten Konsens, dass Verstehen gradierbar ist, möchte ich nun vorschlagen, dass instrumentelles Verstehen als *nicht* gradierbar aufgefasst werden sollte. Der Grund dafür ist der theoretische Zweck meines Explikats (Abschnitt 4.2.3). Ich möchte dasjenige Verstehen charakterisieren, das die epistemische und mitunter auch substanzielle Erfüllung gesellschaftlicher Desiderata ermöglicht. Entweder ein Stakeholder versteht systembezogene Aspekte zu einem ausreichenden Grad, um beurteilen zu können, ob ein gegebenes Desideratum (un)erfüllt ist, oder er tut es nicht. Wenn er weniger versteht als nötig, besitzt er kein instrumentelles Verstehen. Wenn er mehr als das versteht, verbessert das sein instrumentelles Verstehen nicht. Das gilt unabhängig davon, dass bestimmte gesellschaftliche Desiderata zu einem höheren oder geringeren Grad erfüllt sein können: In diesem Fall erfordert instrumentelles Verstehen epistemischen Zugang dazu, ob ein gegebenes Desideratum zu einem kontextuell ausreichenden Grad erfüllt ist. Auch dann ist instrumentelles Verstehen von binärer Qualität.

Die Unterscheidung verschiedener Verstehensgrade und Dimensionen der Gradierbarkeit ist dennoch wichtig, um mein Explikat zu charakterisieren. Denn instrumentelles Verstehen entspricht einem gewissen, kontextuell ausreichenden (hohen oder niedrigen) Verstehensgrad.⁵⁹ Dafür muss klar sein, wodurch sich Verstehensgrade voneinander unterscheiden. Mein Vorschlag lautet, dass Grade von Verstehen *simpliciter* entlang der in Abbildung 5.1 zusammengestellten Dimensionen variieren können – also gemäß der Vollständigkeit des Verstehens, der Faktizität der Verstehensüberzeugungen, dem Grad der Rechtfertigung und des Begreifens relevanter Zusammenhänge, einschließlich der zugehörigen Subdimensionen.

Die folgenden Kapitel knüpfen daher an diese Unterscheidungen an. In Kapitel 6 arbeite ich heraus, wodurch sich Objekte des Verstehens auszeichnen. Um dafür in Frage zu kommen, so mein Vorschlag, müssen Verstehensobjekte komplex genug sein, um das Begreifen von Zusammenhängen zu ermöglichen. In Kapitel 8 bestimme ich den für instrumentelles Verstehen erforderlichen Grad der Faktizität. In Kapitel 9 komme ich auf den Zusammenhang zwischen den beiden Subdimensionen von Begreifen zurück und identifiziere die für instrumentelles Verstehen erforderlichen verstehensbezogenen Fähigkeiten. Schließlich präzisiere ich in Kapitel 10 den nötigen Grad der Rechtfertigung. Wie in den zugehörigen Kapiteln deutlich werden wird, wirken sich jeweils verschiedene Kontextfaktoren

⁵⁹Diese Auffassung ähnelt bestehenden Vorschlägen, den Schwellwert für absolute Zuschreibungen von Verstehen von kontextuellen Faktoren abhängig zu machen (Bachmann, 2020, S. 80; Baumberger, 2019, S. 376; Kelp, 2015, S. 3813–3814; Schmidt, 2026; siehe auch Abschnitt 8.3.3).

darauf aus, welcher Verstehensgrad entlang der einzelnen Dimensionen erforderlich für instrumentelles Verstehen ist.

Eine Dimension der Gradierbarkeit, die ich nicht zur Charakterisierung instrumentellen Verstehens heranziehen werde, ist die Vollständigkeit samt der Subdimensionen Breite, Tiefe und Relevanz. Wie in Abschnitt 5.1.5 deutlich wurde, bilden diese Subdimensionen eher intuitive Unterschiede ab, die sich nicht klar voneinander abgrenzen lassen.⁶⁰ Womöglich ließen sich die einzelnen Subdimensionen der Vollständigkeit präziser definieren und dadurch klarer voneinander abgrenzen. Das hier vertretene intuitive Verständnis bietet dafür aber keine gute Grundlage. Die Definition instrumentellen Verstehens darauf zu stützen, würde vermutlich mehr Probleme aufwerfen als lösen, weil oft uneindeutig wäre, ob eine Proposition z. B. zu einem tieferen oder breiteren Verständnis beiträgt.

Vielleicht ist ja zumindest die übergeordnete Dimension der Vollständigkeit ein relevanter Faktor für instrumentelles Verstehen? Schließlich scheint ein Zusammenhang zu den in Abschnitt 4.2.1 beschriebenen rollenbasierten Expertiseerwartungen zu bestehen – von einer Systementwicklerin erwarten wir beispielsweise ein vollständigeres Verständnis eines KI-Systems als von einer Endnutzerin. Doch soll mein Begriff instrumentellen Verstehens den geringstmöglichen Verstehensgrad beschreiben, der einem Stakeholder die Erfüllung seines Desideratums ermöglicht (auch wenn wir *qua* seiner Rolle legitimerweise ein umfassenderes Verständnis erwarten können, das über die Erfüllung des Desideratums hinausgeht). Sein Verständnis muss nur so vollständig sein, wie es für diesen konkreten Zweck nötig ist. Was das bedeutet, variiert wiederum mit dem Kontext.

Ich wähle daher eine andere, direktere Strategie, ohne einen Umweg über die problembehaftete Dimension der Vollständigkeit zu nehmen: Entscheidend ist, dass ein Stakeholder (wahre) Überzeugungen über kontextuell relevante systembezogene Aspekte besitzt – unabhängig davon, ob das resultierende Verstehen eher breit oder eng, tief oder oberflächlich, vollständig oder lückenhaft ist. Anstatt den für instrumentelles Verstehen erforderlichen Grad der Vollständigkeit festzulegen, werde ich daher im nun folgenden Kapitel die Objekte instrumentellen Verstehens näher bestimmen.

⁶⁰Auffällig ist, dass auch die meisten Verstehenstheoretiker:innen wenig Konkretes dazu sagen. Eine Ausnahme ist Christoph Kelp, der Breite und Tiefe jedoch auf andere der hier genannten Subdimensionen zurückführt: Er scheint anzunehmen, dass die Vollständigkeit der Breite des Verstehens entspricht und die Verknüpftheit der Tiefe (Kelp, 2015, Fn. 13 auf S. 3812; Kelp, 2017, Endnote 4 auf S. 267). In einer späteren Formulierung seines Vorschlags sympathisiert er mit der Position, dass sowohl Vollständigkeit als auch Verknüpftheit die Tiefe des Verstehens beschreiben (Kelp, 2021, S. 110).

6 Objekte des Verstehens

Menschen verstehen die verschiedensten Dinge – Waschmaschinen, radioaktive Strahlung, den Neid des Nachbarn oder *Tetris*. Anders ausgedrückt: Verstehen kann eine Vielzahl unterschiedlicher Objekte haben. Das lässt sich anhand gängiger Verwendungsweisen des Verbs ‚verstehen‘ verdeutlichen:

- | | |
|--|--|
| (1) ‚Fritz versteht Magda.‘ (Menschen) | (9) ‚Sina versteht die Evolutionstheorie.‘ (Theorien) |
| (2) ‚Lenny versteht, warum Julie die Party so früh verlassen hat.‘ (Handlungen) | (10) ‚Anna versteht Ottomotoren.‘ (Artefakte) |
| (3) ‚Rosi versteht Martins Angst vor Spinnen.‘ (mentale Einstellungen oder Emotionen) | (11) ‚Milo versteht das Balkendiagramm.‘ (Repräsentationen) |
| (4) ‚Wini versteht, warum sich an der Nordsee Ebbe und Flut abwechseln.‘ (natürliche Phänomene) | (12) ‚Skadi versteht Hindi.‘ (Sprachen) |
| (5) ‚Christian versteht, warum die CDU die Bundestagswahl 2025 gewonnen hat.‘ (historische Ereignisse) | (13) ‚Denise versteht ‚Windrad.‘ (Worte) |
| (6) ‚Tobias versteht das Studierendenparlament.‘ (Institutionen) | (14) ‚Matti versteht <i>Aporie</i> .‘ (Begriffe) |
| (7) ‚Andi versteht, wie eine hydraulische Bremse funktioniert.‘ (Mechanismen) | (15) ‚Luca versteht <i>Elfer Raus</i> .‘ (Spiele) |
| (8) ‚Gesa versteht Museologie.‘ (Themengebiete) | (16) ‚Jule versteht Frida Kahlos <i>El sueño (La cama)</i> .‘ (Kunst) |
| | (17) ‚Fritzi versteht Jane Austens <i>Emma</i> .‘ (Fiktionale Erzählungen und Narrative) ⁶¹ |

Viele dieser Verstehensobjekte scheinen im Lichte der Charakterisierung meines Explikandums in Abschnitt 4.2 irrelevant für die vorliegende Fragestellung. Instrumentelles Verstehen betrifft offenbar weder Menschen (1), mentale Einstellungen oder Emotionen (3),

⁶¹Bei dieser Auflistung handelt es sich um eine Kombination und geringfügige Adaption ähnlicher Auflistungen von Stefan Sleeuw (2023, S. 12) und Daniel Wilkenfeld (2013, S. 998) für das englische ‚understand‘, das häufig ähnlich verwendet werden kann wie das deutsche ‚verstehen‘.

historische Ereignisse (5), Sprachen (12), Spiele (15), Kunst (16) oder fiktionale Erzählungen und Narrative (17).

Prima facie scheint instrumentelles Verstehen lediglich Artefakte (10), nämlich KI-Systeme oder einzelne ihrer Aspekte, zum Gegenstand zu haben. Bei genauerer Betrachtung zeigen sich aber weitere Überschneidungen mit der obigen Auflistung: Das linguistische Verstehen von relevanten Worten (13) oder Begriffen (14) ist offensichtlich eine wichtige Voraussetzung, um Aspekte von KI-Systemen zu verstehen (mehr dazu in Abschnitt 9.2.2). Zudem kann auch das Verstehen von bestimmten Themengebieten (8) – z. B. Maschinelles Lernen – zu instrumentellem Verstehen beitragen. In Abschnitt 4.2.1 wurde deutlich, dass ein solches Verständnis Teil jener Expertise sein kann, die wir *qua* seiner Rolle legitimerweise von einem Stakeholder erwarten können.

Weitere Überschneidungen treten zutage, wenn man philosophische Theorien der Erklärung und ihre Anwendbarkeit auf KI-Systeme betrachtet (Abschnitt 3.3): Instrumentelles Verstehen kann dem Verstehen von Handlungen (2) insofern ähneln, dass Gründe-Erklärungen üblicherweise menschliche Handlungen erklären, aber auch auf KI-Systeme angewandt werden können. Instrumentelles Verstehen kann zudem Mechanismen (7) betreffen, insofern mechanistische Erklärungen (wenigstens metaphorisch) genutzt werden können, um KI-Systeme verständlich zu machen. Außerdem können diverse Repräsentationen (11) Gegenstand instrumentellen Verstehens sein, da in XAI-Kontexten z. B. Salienz-karten, linguistische Erklärungen oder Diagramme verwendet werden, um Aspekte von KI-Systemen transparent zu machen.

Weiterhin können spezifische Desiderata (siehe Abschnitt 4.2.3) den Kreis der Objekte instrumentellen Verstehens erweitern: So kann sich das Verstehen natürlicher Phänomene (4) und (wissenschaftlicher) Theorien (9) mit instrumentellem Verstehen überschneiden, weil ein mögliches Desideratum, das mithilfe von XAI erfüllt werden soll, das Erlangen wissenschaftlicher Einsichten mithilfe von KI ist. In ähnlicher Form könnte auch das Verstehen bestimmter Institutionen (6) für die Erfüllung von Desiderata erforderlich sein – etwa, wenn es darum geht, jemanden bei Fehlern zur Rechenschaft zu ziehen.

Diese Auflistung möglicher Verstehensobjekte ist zwangsläufig unvollständig. Das Verb ‚verstehen‘ kann flexibel auf unterschiedlichste Objekte angewandt werden, und selbst die Aspekte von KI-Systemen oder den übergeordneten soziotechnischen Aspekten, deren Verständnis potenziell zu instrumentellem Verstehen beitragen kann, sind vermutlich unzählig. Angesichts dessen ist es attraktiv, allgemeinere Oberkategorien aufzustellen, in die sich verschiedene Verstehensobjekte eingliedern lassen. In der Verstehensliteratur wird in diesem Zusammenhang häufig zwischen propositionalem, interrogativem, objektualem, symbolischem und faktischem Verstehen unterschieden. Allerdings können diese Kategorien unterschiedlich aufgefasst werden und häufig bleibt unklar, welche Lesart jeweils intendiert ist. Dieses Kapitel soll diese Unklarheiten auflösen und für instrumentelles Verstehen relevante Verstehensobjekte identifizieren.

Ich gehe wie folgt vor: In Abschnitt 6.1 diskutiere ich die Unterscheidung zwischen propositionalem, interrogativem und objektualem Verstehen und stelle drei Möglichkei-

ten vor, wie diese Kategorisierung interpretiert werden kann. In Abschnitt 6.2 gehe ich auf symbolisches sowie faktisches Verstehen ein, das meinem Vorschlag zufolge die Repräsentationsebene des Verstehensobjekts betrifft. In Abschnitt 6.3 übertrage ich diese Überlegungen auf mein Explikat. Ich argumentiere zum einen, dass instrumentelles Verstehen häufig gleichermaßen als faktisch und symbolisch aufgefasst werden kann, weil XAI-Kontexte komplexe Repräsentationsbeziehungen involvieren. Zum anderen plädiere ich im Kontext dieser Arbeit dafür, den Unterschied zwischen propositionalem, interrogativem und objektualem Verstehen näherungsweise als Differenz im Komplexitätsgrad des Verstehensobjekts aufzufassen.

6.1 Propositionales, interrogatives und objektuales Verstehen

Es ist üblich, zwischen den folgenden drei Kategorien von Verstehensobjekten zu unterscheiden:

1. *Propositionales Verstehen* oder *Verstehen-dass*. Beispiel: Elif versteht, dass das Expertensystem eine Gastritis diagnostiziert hat.
2. *Interrogatives Verstehen* oder *Verstehen-w*. Beispiele: Elif versteht, warum das Expertensystem eine Gastritis diagnostiziert hat; wann die Diagnose gestellt wurde; wo das Expertensystem entwickelt wurde; wie das System funktioniert; wer der Patient ist; was der Unterschied zwischen einer Typ-A- und Typ-B-Gastritis ist usw.
3. *Objektuales Verstehen*. Beispiele: Elif versteht das Expertensystem; die Symptome einer Gastritis; die Chefentwicklerin des Systems.

Diese drei Oberkategorien von Verstehensobjekten haben unterschiedlich viel Aufmerksamkeit von Philosoph:innen bekommen. *Propositionales Verstehen* wird häufig als erkenntnistheoretisch uninteressant oder abseitig verworfen (z. B. Baumberger, 2011, S. 87; Elgin, 2007, S. 34; Gordon, 2012; Grimm, 2011, Kap. 1; McCain, 2016, Fn. 28 auf S. 146; Pritchard, 2009, S. 30–31). Dahinter steckt meist die Auffassung, dass kein interessanter Unterschied zwischen propositionalem Verstehen und propositionalem Wissen bestehe. Da Letzteres in der Erkenntnistheorie bereits ausgiebig verhandelt werde, so die Annahme, gebe es keinen Bedarf für eine gesonderte Theorie propositionalen Verstehens. Von den unterschiedlichen Varianten *interrogativen Verstehens* wird vornehmlich *Verstehen-warum* bzw. *explanatorisches Verstehen* diskutiert (z. B. Hills, 2016; Khalifa, 2017; Newman, 2017a). Das liegt nicht zuletzt am in Abschnitt 3.3.2 geschilderten Zusammenhang zwischen Verstehen und Erklärungen, der explanatorisches Verstehen gleichermaßen für Erkenntnis- und Wissenschaftstheoretiker:innen interessant macht. Auch *objektuales Verstehen* erfährt viel Aufmerksamkeit (z. B. Baumberger, 2019; Carter & Gordon, 2016; Elgin, 2017; Kvanvig, 2003). Ein wichtiger Grund dafür dürfte sein, dass Themengebiete, Artefakte oder andere umfangreiche, komplexe Dinge typische Verstehensobjekte darstellen.

Neben der Analyse dieser Verstehensobjekte existiert eine Subdebatte dazu, ob diese Oberkategorien jeweils distinkte Verstehensarten konstituieren bzw. ob sich manche dieser Verstehensarten auf andere reduzieren lassen. Die Frage ist also, ob bestimmte, anhand des Objekts individuierte Verstehensarten metaphysisch fundamentaler sind als andere, bzw. ob ein bestimmtes Verstehensobjekt in Form eines anderen erklärt werden kann. Diskutiert wird insbesondere, ob objektuales Verstehen auf explanatorisches (z. B. Baumberger, 2011; Dellsén, 2020; Grimm, 2017, Kap. 3; Khalifa, 2017, Kap. 4) oder propositionales (etwa Kvanvig, 2009b, Kap. 2) Verstehen reduziert werden kann und ob explanatorisches auf propositionales Verstehen reduzierbar ist (z. B. Baumberger, 2011, Kap. 5; Egler, 2021, S. 6074–6075; Kvanvig, 2003, S. 189–190).

Diese Debatten werden dadurch verkompliziert, dass nicht immer klar ist, wodurch verschiedene Autor:innen die besagten Verstehensobjekte voneinander abgrenzen. Bevor ich für instrumentelles Verstehen relevante Objekte und ihr Verhältnis untereinander bestimmen kann, müssen wir daher klären, was genau den Unterschied zwischen propositionalem, interrogativem und objektualem Verstehen ausmacht. Darauf gibt es meines Erachtens drei Antworten: Verschiedene Verstehensobjekte sind unterschiedliche grammatikalische Objekte (Abschnitt 6.1.1), Verstehensobjekte sind entweder Propositionen oder Dinge (Abschnitt 6.1.2), oder Verstehensobjekte unterscheiden sich durch ihre Komplexität (Abschnitt 6.1.3).

6.1.1 Verstehensobjekte als grammatikalische Objekte

Die erste Möglichkeit, die Unterscheidung zwischen propositionalem, interrogativem und objektualem Verstehen auszubuchstabieren, ist die Differenzierung anhand verschiedener *grammatikalischer* Objekte. Das Objekt objektualen Verstehens ist demnach ein Nomen oder eine Nominalphrase – wie ‚Expertensystem‘ oder ‚die Symptome einer Gastritis‘. Das Objekt interrogativen Verstehens ist die Antwort auf einen Fragesatz, der durch ein Interrogativpronomen bzw. -adverb eingeleitet wird (und vielleicht auch der Fragesatz selbst).⁶² Zum Beispiel beantwortet ‚weil eine Helicobacter-Infektion detektiert wurde‘ die Frage danach, ‚warum das Expertensystem eine Gastritis diagnostiziert hat‘. Schließlich ist das Objekt propositionalen Verstehens ein Aussagesatz, der mit ‚dass‘ beginnt – wie beispielsweise ‚dass das Expertensystem eine Gastritis diagnostiziert hat‘.

Aus epistemologischer Sicht handelt es sich dabei aber nur um oberflächliche Unterschiede. Die Satzstruktur kann in vielen Fällen abgewandelt werden, ohne dass sich die Bedeutung maßgeblich verändert (vgl. Baumberger, 2011, S. 71; Bengson, 2018, S. 15–16; Sleeuw, 2023, S. 10–11):

- Elif versteht, warum das Expertensystem eine Gastritis diagnostiziert hat.
- Elif versteht, wie die Diagnose des Systems zustande gekommen ist.

⁶²Eine Übersicht über diese Debatte in Bezug auf interrogatives Wissen bietet Berit Brogaard (2009).

- Elif versteht die Ursache der Diagnose des Expertensystems.
- Elif versteht, dass die Diagnose des Systems durch die Detektion einer Helicobacter-Infektion verursacht wurde.

Die obigen Sätze beschreiben offenbar dasselbe Verstehensobjekt, obwohl das grammatikalische Objekt wechselt. Umgekehrt scheinen die folgenden Sätze, die allesamt einen Interrogativsatz als grammatikalisches Verstehensobjekt ausweisen, aus erkenntnistheoretischer Perspektive unterschiedliche Verstehensobjekte zu beschreiben (vgl. ebd.):

- Elif versteht, warum das Expertensystem eine Gastritis diagnostiziert hat.
- Elif versteht, wo sich das System befindet.
- Elif versteht, wann das Expertensystem entwickelt wurde.

Diese Beispiele verdeutlichen zweierlei. Erstens: Die Antwort auf die Frage, ob bestimmte Verstehensobjekte auf andere reduziert werden können, muss in Bezug auf *grammatikalische* Objekte ‚Ja‘ lauten. Zwar muss man dafür womöglich eine umständliche Formulierung in Kauf nehmen, doch lassen sich objektuales und interrogatives Verstehen aus linguistischer Sicht auf propositionales Verstehen reduzieren. Das objektuale Verstehen eines Expertensystems ließe sich auch als (lange) Konjunktion von propositionalem Verstehen (in Form von Aussagen über die Wissensbasis, Erklärungskomponenten, Wenn-dann-Regeln usw.) ausdrücken; das interrogative Verstehen einer Gastritis-Diagnose durch eine Konjunktion von Aussagen über Infektionen, Symptome, Risikofaktoren etc.

Zweitens: Aus linguistischer Perspektive haben grammatikalische Konstruktionen, die ein interrogatives oder objektuales Verstehensobjekt aufrufen, zwar Vorteile. Beispielsweise muss man nicht benennen, warum das Expertensysteme eine Diagnose getroffen hat, um Elif explanatorisches Verstehen zuzuschreiben, und man kann vom objektualen Verstehen eines großen Themengebiets (z. B. Informatik) sprechen, ohne sämtliche, potenziell veränderliche Propositionen auflisten zu müssen, die dem Thema angehören (vgl. Farkas, 2019, S. 266; Khalifa, 2017, S. 124). Die grammatikalischen Objektkategorien sind aber zu unbeständig, um epistemologisch interessante Verstehensobjekte voneinander zu unterscheiden. Für meine Zwecke ist es daher nicht hilfreich, Verstehensobjekte anhand dieser Kategorien voneinander abzugrenzen.

6.1.2 Verstehensobjekte als Propositionen oder Dinge

Die meisten Erkenntnistheoretiker:innen scheinen bei der Unterscheidung zwischen propositionalem, interrogativem und objektualen Verstehen eine ähnliche Unterscheidung aus der Wissensdebatte im Sinn zu haben. In Bezug auf Wissen ist es üblich, drei Arten zu unterscheiden (z. B. Feldman, 2003, S. 12):⁶³

⁶³Ich folge hier im Wesentlichen Katalin Farkas' Darstellung dieser drei Wissensarten (Farkas, 2019).

1. *Propositionales Wissen* oder *Wissen-dass*. Beispiel: Elif weiß, dass das Expertensystem eine Gastritis diagnostiziert hat.
2. *Praktisches Wissen* oder *Wissen-wie*. Beispiel: Elif weiß, wie man ein Expertensystem programmiert.
3. *Objektuales Wissen* oder *Wissen durch Bekanntschaft* (engl. *knowledge by acquaintance*). Beispiel: Elif kennt⁶⁴ die Chefentwicklerin des Systems.

Der mit Abstand größte Teil der philosophischen Debatte zu Wissen befasst sich mit *propositionalem Wissen*. Propositionales Wissen erfordert eine mentale Einstellung (üblicherweise wahre Überzeugung) in Bezug auf etwas, das wahr oder falsch sein kann (üblicherweise eine Proposition), klassischerweise ergänzt um zusätzliche Bedingungen wie Rechtfertigung oder eine Anti-Zufalls-Klausel. *Interrogatives Wissen* (Wissen-wo, -wann, -wer...) gilt gemeinhin als Spezialfall propositionalen Wissens: Interrogatives Wissen ist einfach Wissen der Proposition, welche die jeweilige W-Frage beantwortet (und ggf. auch der W-Frage selbst).

Darüber hinaus gibt es eine kontroverse Subdebatte zu *praktischem Wissen*. Der intellektualistischen Auffassung zufolge ist praktisches Wissen ebenfalls eine Form propositionalen Wissens (Stanley & Williamson, 2001). Anti-Intellektualist:innen nehmen hingegen an, dass es sich dabei eher um die Fähigkeit oder Disposition handelt, eine bestimmte Handlung erfolgreich auszuführen (z. B. Ryle, 1949).

Objektuales Wissen bzw. *Wissen durch Bekanntschaft* erfährt im Vergleich dazu nur wenig philosophische Aufmerksamkeit (Duncan, 2023). Zurückgeführt wird diese Wissensart zumeist auf Bertrand Russell (1910), der Wissen durch Bekanntschaft als eine Form des *Wissens von Dingen* charakterisiert und von dem (propositionalen) *Wissen von Wahrheiten* abgrenzt. Wissen durch Bekanntschaft hat man nur, wenn man das fragliche Objekt unmittelbar erfährt. Beispielsweise weiß ich (propositional), dass Manfred Bayer der Rektor der TU Dortmund ist, aber ich kenne ihn nicht (wir sind uns nie begegnet). Umgekehrt kann ich Objekte oder Menschen kennen (erfahren), ohne etwas (propositional) über sie zu wissen. Wissen durch Bekanntschaft wird in der aktuellen Debatte meist mit objektualem Wissen gleichgesetzt (Farkas, 2019, S. 262–263). Bei objektualem Wissen handelt es sich also nicht um eine Einstellung gegenüber einer Proposition, sondern es besteht in bzw. erwächst aus einer bestimmten Erfahrung. Das macht sowohl das Objekt objektualen Wissens (ein Ding in der Welt) als auch die epistemische Einstellung gegenüber diesem Objekt (eine phänomenale Erfahrung) nicht-propositional.

Auch wenn diese klassische Unterscheidung zwischen Wissensarten von Teilen der Verstehensdebatte übernommen wurde, setzen Erkenntnistheoretiker:innen in Bezug auf Verstehen meist andere Schwerpunkte bzw. legen die Unterscheidungen anders aus, als

⁶⁴Im Deutschen werden Konstruktionen, die Wissen durch Bekanntschaft ausdrücken, eher mit ‚kennen‘ denn ‚wissen‘ gebildet; dasselbe Phänomen gibt es auch in anderen Sprachen wie Französisch oder Ungarisch (Farkas, 2019, S. 264).

es bezüglich Wissen üblich ist. Die erste Diskrepanz besteht darin, dass innerhalb der Verstehensdebatte *interrogatives Verstehen* oft gesondert hervorgehoben wird, während die analoge Kategorie in Bezug auf Wissen üblicherweise in der Oberkategorie des propositionalen Wissens aufgeht. Ein wichtiger Grund dafür ist, wie oben erwähnt, dass Verstehen intuitiv eng mit Erklärungen zusammenhängt und daher viele Philosoph:innen interrogativem (insbesondere: explanatorischem) Verstehen gesonderte Aufmerksamkeit schenken. Ein weiterer Grund besteht darin, dass die Reduzierbarkeit interrogativen Verstehens auf propositionales Verstehen kontroverser ist als die analoge Position in der Wissensdebatte. Zwar nehmen auch in Bezug auf Verstehen viele Autor:innen an, dass die interrogative Variante letztlich auf die propositionale Variante reduziert werden kann, allerdings haben andere Philosoph:innen dieser Position zumindest in Bezug auf explanatorisches Verstehen widersprochen.⁶⁵ Die Argumentation für letztere Position stützt sich häufig auf die verbreitete Annahme, dass propositionales Verstehen auf propositionales Wissen reduziert werden könne. Explanatorisches Verstehen erfordere hingegen mehr als das, z. B. das Begreifen von Zusammenhängen und/oder distinkte Fähigkeiten. Daher sei eine verlustfreie Reduktion auf propositionales Verstehen nicht möglich. Ich komme in Abschnitt 6.3 auf diese Argumentation zurück.

Eine zweite Diskrepanz besteht darin, dass Wissen-wie zwar kontrovers diskutiert wird, die Debatte um *Verstehen-wie* allerdings ein Nischendasein fristet. Die meisten Autor:innen übergehen das Thema vollkommen, manche machen beiläufige Bemerkungen (etwa Gijsbers, 2013, S. 518; Khalifa, 2017, S. 2; Kvanvig, 2003, S. 190; P. Lipton, 2009, S. 54). Dezidierte Auseinandersetzungen mit praktischem Verstehen sind hingegen selten. Ich möchte kurz auf zwei hervorhebenswerte Ausnahmen eingehen (aber siehe auch Boyd, 2017, und Hannon, 2021, für einige Überlegungen zum Thema). Die erste dieser Ausnahmen ist ein Aufsatz von John Bengson. Bengson, der an seine Arbeiten zu Wissen-wie anknüpft, argumentiert für die ‚Einheit‘ von praktischem und theoretischem Verstehen (Bengson, 2018). Seinem Vorschlag zufolge sind sowohl theoretisches als auch praktisches Verstehen etwas Intellektuelles (anstatt z. B. etwas Behavioral-dispositionales). Theoretisches Verstehen sei erhellend (engl. *illuminating*), praktisches Verstehen sei handlungsleitend (engl. *action-guiding*); doch es handle sich nicht um distinkte Verstehensarten.

Die zweite Ausnahme ist ein Beitrag von Michael Stuart, der sich mit KI-gestützter Wissenschaft befasst. Stuart definiert praktisches – in seiner Terminologie: pragmatisches – Verstehen relativ zu einem (z. B. KI- oder natürlichen) System durch zwei notwendige und zusammen hinreichende Bedingungen: Die erste verlangt das Verfügen über die Fertigkeiten,⁶⁶ das System oder seine Bestandteile robust und erfolgreich derart zu manipulieren,

⁶⁵Eine reduktive Position vertreten etwa Boyd, 2017, S. 110; Hannon, 2021, S. 283; Kvanvig, 2003, S. 189; Moravcsik, 1979 und Strevens, 2013. Für eine antireduktionistische Auffassung argumentieren unter anderem Baumberger, 2011, Kap. 5; Egler, 2021, S. 6074–6075 sowie Grimm, 2011, S. 87–88.

⁶⁶Stuart grenzt Fertigkeiten (engl. *skills*) von Fähigkeiten (engl. *abilities*) ab: Im Gegensatz zu Fähigkeiten seien Fertigkeiten aufgabenspezifisch, sie würden intentional erworben und angewandt und unterlägen der kognitiven Kontrolle des Subjekts.

dass die eigenen Ziele erreicht werden. Die zweite Bedingung erfordert, dass das verstehende Subjekt dafür verantwortlich bzw. lobenswert ist, dass es über diese Fertigkeiten verfügt (Stuart, 2025). Stuart schlägt außerdem vor, dass pragmatisches Verstehen in manchen Fällen notwendig für (nicht-pragmatisches) explanatorisches oder objektuales Verstehen sein kann.

Abgesehen von Bengson und Stuart haben sich meines Wissens nur wenige Autor:innen mit Verstehen-wie befasst. Vermutlich wird praktisches Verstehen deswegen weniger diskutiert, weil Verstehen jeglicher Form ohnehin Ähnlichkeit mit Wissen-wie aufweist. Denn wer etwas versteht, besitzt üblicherweise charakteristische Fähigkeiten. Wir werden uns in Kapitel 9 ausführlich mit diesem Aspekt von Verstehen befassen.

Neben diesen Unterschieden gibt es auch Überschneidungen zwischen den Debatten zu Objektkategorien von Wissen bzw. Verstehen. Wie oben dargestellt, entspricht objektuales Wissen der Standardauffassung zufolge dem nicht-propositionalen Wissen, das man durch Bekanntheit bzw. direkte Erfahrung erwirbt. Manche Aspekte dieser Wissensform finden sich auch in der Verstehensdebatte wieder. Beispielsweise haben einige Philosoph:innen vorgeschlagen, dass sich Verstehen nicht zwangsläufig auf propositionale Einstellungen reduzieren lasse. So argumentiert Julius Moravcsik, dass objektuales Verstehen mit einer kognitiven, nicht-propositionalen Repräsentation des Verstandenen einhergehe, aufgrund derer das verstehende Subjekt über spezifische Fähigkeiten verfüge. Dieses ‚Begreifen‘ bzw. diese ‚Einsicht‘ mache objektuales Verstehen nicht-propositional (Moravcsik, 1979). In ähnlicher Form definiert Linda Zagzebski Verstehen als „the state of comprehension of nonpropositional structures of reality“ (Zagzebski, 2001, S. 242; Hervorhebung entfernt). Weder sei die Realität notwendigerweise propositional strukturiert, noch seien die mentalen Repräsentationen dieser Realität zwangsläufig propositional (z. B. könnten sie auch nicht-propositionale Karten, Modelle oder Diagramme beinhalten; ebd., S. 241). Andere haben argumentiert, dass Verstehen bzw. Begreifen phänomenales Erleben erfordert. Wer z. B. wirklich verstehe, dass das Volumen der Sonne 1,3 Millionen Mal größer ist als das der Erde, oder dass Fleischkonsum zu Tierleid führt, habe auch eine phänomenale Erfahrung dieser Dinge bzw. der zugehörigen Propositionen (die Beispiele stammen von Bourget, 2024).

Ich komme im weiteren Verlauf der Arbeit auf diese nicht-propositionalen Aspekte von Verstehen zurück. An dieser Stelle möchte ich lediglich festhalten, dass eine Unterteilung von Verstehensobjekten in Propositionen und nicht-propositionale Strukturen bzw. Erfahrungen – anders als die Differenzierung grammatikalischer Objekte – *prima facie* dafür spricht, dass unterschiedliche Verstehensobjekte verschiedene, irreduzible Verstehensarten markieren.

6.1.3 Verstehensobjekte und Komplexitätsgrade

Eine dritte Möglichkeit der Unterscheidung von Verstehensobjekten lautet, dass sich die jeweiligen Objekte propositionalen, interrogativen und objektualen Verstehens durch ihren

Umfang oder ihre Komplexität unterscheiden. Diese Auffassung ist kompatibel damit, dass auch interrogative und objektuale Verstehensobjekte propositional sind.⁶⁷ Das objektuale Verstehen eines Themengebiets (z. B. ‚Informatik‘) umfasst demnach eine größere Anzahl an (adäquat verknüpften) Propositionen als das explanatorische Verstehen, warum ein medizinisches Expertensystem eine bestimmte Diagnose getroffen hat. Das propositionale Verstehen, dass das System eine Gastritis diagnostiziert hat, würde im Vergleich dazu die kleinste, am wenigsten komplexe Menge an Propositionen umfassen.

Doch wieso sollten auch die Objekte propositionalen Verstehens komplex sein, anstatt eine einzelne Proposition zu umfassen? Der Grund ist der holistische Charakter von Verstehen. Es herrscht weitestgehender Konsens darüber, dass Verstehen eng mit dem Begreifen oder Erfassen von Zusammenhängen verknüpft ist. Eine Person, die versteht, hat nicht bloß voneinander unabhängige Überzeugungen über individuelle Propositionen, sondern begreift auch, wie diese Propositionen miteinander zusammenhängen und vielleicht auch, warum diese Zusammenhänge bestehen.⁶⁸ Hier ist ein Beispiel von Stephen Grimm, anhand dessen sich die Intuition verdeutlichen lässt, dass Verstehen auch Begreifen erfordert:

Volkswagen: „Suppose that one day I get into my 1991 Volkswagen and none of the gauges—the speedometer, tachometer, etc.—come to life; they’re all dead. I drive the car to the garage and my mechanic tells me that the reason why the gauges are dead is because I have a bad ignition switch. I then seem to have excellent reason to assent to the claim that my gauges are dead because of a bad ignition switch, even though I fail to grasp how a bad ignition switch might lead to this result. [...] As this discussion suggests, when trying to offer an account of understanding the notion of *grasping* arises almost irresistibly.“ (Grimm, 2006, S. 531–532; Hervorhebung im Original)

Intuitiv versteht Stephen Grimms Alter Ego in *Volkswagen* nicht, warum die Kontrollinstrumente seines Fahrzeugs ausgefallen sind, weil er nicht ‚sieht‘ oder ‚begreift‘ (engl. *seeing* bzw. *grasping*), wie die defekte Zündung mit diesem Totalausfall zusammenhängt.

⁶⁷Dasselbe gilt auch für viele Formen von Wissen, deren (grammatikalisches) Objekt ein Nomen ist und die im grammatikalischen Sinne daher objektuales Wissen darstellen (siehe Farkas, 2019, Kap. 5).

⁶⁸Entsprechende Bemerkungen sind allgegenwärtig. Einige Beispiele: „Understanding why p [...] requires more than the correct belief that p because q. It requires a grasp of the reason why p, or more precisely, a grasp of the relationship between p and q“ (Hills, 2016, S. 663). „[...] understanding [...] is the appreciation or grasp of order, pattern, and how things ‘hang together’“ (Riggs, 2003, S. 217). „Understanding requires the grasping of explanatory and other coherence-making relationships in a large and comprehensive body of information. One can know many unrelated pieces of information, but understanding is achieved only when informational items are pieced together by the subject in question“ (Kvanvig, 2003, S. 192). „The discernment of patterns, of similarities and differences, of connections and disconnections, is essential to understanding“ (Cooper, 1995, S. 206). „[...] understanding is not directed toward a discrete object, but involves seeing the relation of parts to other parts and perhaps even the relation of part to a whole. It follows that the object of understanding is not a discrete proposition“ (Zagzebski, 2001, S. 241). „Understanding [...] is entirely a matter of fitting into a pattern. Understanding depends on coherence“ (Kosso, 2007, S. 181).

Ein solches Begreifen scheint notwendig für Verstehen zu sein. Die potenziell begriffenen Zusammenhänge sind dabei so vielfältig wie die möglichen Verstehensobjekte. Neben den klassischen kohärenzstiftenden Zusammenhängen – logisch, probabilistisch, explanatorisch – haben verschiedene Autor:innen unter anderem teleologische, definitorische, analogiebildende, reduktiv/vereinigende, kausale, modale, mereologische, raum-zeitliche, ähnlichkeitsbasierte, supervenierende, begriffliche und semantische Zusammenhänge vorgeschlagen (Baumberger, 2011, S. 79; Gijsbers, 2013; Greco, 2014, S. 293; Kvanvig, 2009b, S. 96–97; M. P. Lynch, 2018, S. 197; van Camp, 2014, S. 105; natürlich bestehen teils Überlappungen zwischen diesen Zusammenhängen).⁶⁹ Welche Zusammenhänge begriffen werden müssen, um zu verstehen, hängt davon ab, welche Zusammenhänge in der ein oder anderen Form im jeweiligen Objekt des Verstehens vorliegen. Wie wir in Abschnitt 5.1.5 festgestellt haben, betrifft dies auch die Faktizitätsdimension von Verstehen: Um akkurat bzw. faktiv zu sein, muss das Überzeugungssystem des verstehenden Subjekts die tatsächlich existierenden Zusammenhänge akkurat widerspiegeln.

Wenn das Begreifen von Zusammenhängen notwendig für Verstehen ist, zeichnen sich Verstehensobjekte zwangsläufig durch eine gewisse Komplexität aus – eine isolierte, simple Proposition beinhaltet schließlich keine Zusammenhänge, die begriffen werden könnten. In diesem Sinne ist Verstehen niemals atomistisch, sondern stets holistisch.⁷⁰ Folglich ist auch propositionales Verstehen nicht auf die Proposition begrenzt, die durch den zugehörigen ‚dass‘-Satz ausgedrückt wird. Es erfordert darüber hinaus ein komplexeres Verständnis relevanter Zusammenhänge, in welche die fragliche Proposition eingebettet ist. Wie Catherine Elgin es ausdrückt:

If the proposition ‘I understand that Athens defeated Persia in the battle of Marathon’ is supposed to be a stand-alone proposition [...], it is hard to see how it differs from ‘I know that Athens defeated Persia in the battle of Marathon’. [...] But if my understanding that Athens defeated Persia in the battle of Marathon depends in a suitable way on my overall understanding of such matters as the course of the battle, the strategies and tactics of the two armies, and the history of relations between Athens and Persia, the situation is different. The epistemological standing of ‘Athens defeated Persia in the battle of Marathon’ then derives from its place in a more comprehensive understanding of the history of Greece. [...] The understanding encapsulated in individual propositions derives from an understanding of larger bodies of information that include those propositions. (Elgin, 2007, S. 34–35)

Dieser Lesart zufolge stellen propositionales, interrogatives und objektuales Verstehen keine distinkten Verstehensarten dar, sondern bilden näherungsweise graduelle Unterschiede

⁶⁹Einige dieser Zusammenhänge haben wir bereits in Abschnitt 3.3.2 kennen gelernt, als es um Verstehen ohne Erklärungen ging. Dass nicht alle kohärenzstiftenden Zusammenhänge *explanatorische* Zusammenhänge sind (BonJour, 1985, S. 100), spricht ebenfalls dafür, dass Verstehen auch ohne Erklärungen erworben werden kann.

⁷⁰Die Terminologie ist entlehnt von Catherine Elgin (z. B. 2007) und Duncan Pritchard. Pritchard (z. B. 2010) bezeichnet lediglich objektuales Verstehen als ‚holistisch‘ und explanatorisches Verstehen als ‚atomistisch‘.

de in Umfang und Komplexität des Verstehensobjekts ab. Ähnliche Überlegungen finden sich bei unterschiedlichen Autor:innen wieder. Beispielsweise schlägt Kareem Khalifa vor, dass vorgebliche Fälle von objektuaem Verstehen verlustfrei durch mehrere Instanzen explanatorischen Verstehens beschrieben werden können; ggf. ergänzt um ‚Proto-Verstehen‘, das eine Vorstufe explanatorischen Verstehens darstellt. Objektuaes Verstehen sei daher eine letztlich überflüssige Kategorie (Khalifa, 2017, Kap. 4). Stephen Grimm konstatiert, dass die Unterschiede zwischen explanatorischem und objektuaem Verstehen häufig überbewertet würden. Beide Verstehensformen erwachsen aus dem Begreifen von Abhängigkeitsbeziehungen – bei komplexeren (z. B. objektualen) Verstehensobjekten seien diese Abhängigkeitsbeziehungen eben umfangreicher (Grimm, 2017, Kap. 3; dem schließt sich auch Pérez, 2019, S. 453, an). Und neben Elgin und mir haben auch diverse andere Autor:innen eine holistische Auffassung propositionalen Verstehens befürwortet oder zumindest in Betracht gezogen (Gordon, 2012, Kap. 5; Kvanvig, 2003, S. 192; Malfatti, 2023, S. 20; Zagzebski, 1996, S. 49).

Eine interessante Konsequenz dieser Position ist, dass – entgegen der Standardauffassung – demnach auch propositionales Verstehen gradierbar ist (siehe Kapitel 5): Elif kann demzufolge beispielsweise besser oder schlechter verstehen, dass das Expertensystem eine Gastritis diagnostiziert hat, weil sie mehr oder weniger Zusammenhänge begreifen kann, die diese Proposition betreffen.

Unterscheidet man Verstehensobjekte anhand ihrer Komplexität, ist der gemeinsame Nenner von propositionalem, interrogativem und objektuaem Verstehen, dass ihre Objekte komplex genug sind, um das für Verstehen charakteristische Begreifen von Zusammenhängen zu ermöglichen. Gleichzeitig sind dieser Lesart zufolge interrogatives und objektuaes Verstehen letztlich auf propositionales Verstehen verschiedenen Umfangs reduzierbar – zumindest auf der inhaltlichen Ebene. Durch das in Verstehen involvierte Begreifen von Zusammenhängen, und die Verknüpftheit der verstandenen Propositionen, kommt ein zusätzliches Element ins Spiel, das sich nicht zwangsläufig auf propositionale Einstellungen reduzieren lässt. Ich komme in Abschnitt 6.3 auf diesen Aspekt zurück.

6.2 Symbolisches und faktisches Verstehen

Neben 1. propositionalem, 2. interrogativem und 3. objektuaem Verstehen wird häufig auch zwischen faktischem und symbolischem Verstehen unterschieden:

4. *Faktisches Verstehen*. Beispiele: Elif versteht, dass das Expertensystem eine Gastritis diagnostiziert hat, warum diese Diagnose getroffen wurde; die Architektur künstlicher neuronaler Netzwerke.
5. *Symbolisches Verstehen*. Beispiele: Elif versteht die Visualisierung der Netzwerkstruktur neuronaler Netze; die Erklärung der Gastritis-Diagnose; den Satz ‚Das Expertensystem ist veraltet‘.

Faktisches Verstehen hat als Objekt ein Faktum, Ding oder Ereignis. Jemand mit faktischem Verstehen versteht also einen Aspekt der Welt. Symbolisches Verstehen betrifft hingegen eine symbolische Repräsentation, z. B. natürlichsprachliche Sätze, Erklärungen, Theorien, Narrative, Abbildungen oder Diagramme. Die Unterscheidung betrifft gewissermaßen die Repräsentationsebene des Verstehensobjekts: Handelt es sich um eine Repräsentation, lässt es sich symbolisch verstehen. Das Objekt faktischen Verstehens ist hingegen ein Ding in der Welt, das potenziell als Repräsentandum in Frage kommt.⁷¹

Bei genauerer Betrachtung ist diese Unterscheidung allerdings nicht immer trennscharf. Ein Beispiel dafür sind naturwissenschaftliche Disziplinen oder Themengebiete (Sleeuw, 2023, S. 14). Ist z. B. die Hydrogeographie ein symbolisches oder faktisches Verstehensobjekt? Einerseits befasst sich die Hydrogeographie mit Aspekten der Welt – mit Flüssen, Gletschern oder Grundwasserströmen. Andererseits bedient sie sich eines speziellen Symbolsystems, um diese natürlichen Phänomene zu erforschen. Ich denke, Stefan Sleeuw hat Recht damit, dass beide Lesarten valide sind und je nach Kontext die eine oder die andere im Vordergrund stehen kann (ebd.). In Abschnitt 6.3 werde ich auf diese Problematik zurückkommen, da in XAI-Kontexten mindestens ebenso ambige Beispiele existieren.

Wie verhalten sich symbolisches bzw. faktisches Verstehen einerseits zu propositionalem, interrogativem bzw. objektualem Verstehen andererseits? Im Kontrast zur Systematisierung von z. B. Malfatti (2023, S. 18–19) oder Sleeuw (2023, S. 13–14) finde ich es hilfreich, die beiden Unterscheidungen als orthogonal zueinander aufzufassen. Zwar ist das Objekt symbolischen Verstehens ein Symbol oder eine sonstige Repräsentation und in diesem Sinne stets objektiv. Doch wenn man der in Abschnitt 6.1.3 vorgestellten Auffassung von Verstehensobjekten folgt, ist der Unterschied zwischen propositionalem, interrogativem und objektualem Verstehen fließend. Demnach können Symbole, Repräsentationen, Narrative oder andere Objekte symbolischen Verstehens plausiblerweise auch propositionales oder interrogatives Verstehen *im Kontext des zugehörigen Symbolsystems* vermitteln.

Die Idee lässt sich besonders gut anhand bekanntermaßen falscher Theorien verdeutlichen: Man kann *symbolisch* das Themengebiet Astrologie verstehen (objektiv), oder *dass* eine im Tierkreiszeichen ‚Wassermann‘ geborene Person der Astrologie zufolge rebellisch ist (propositional), oder *warum* sie aus astrologischer Sicht gerne verweist (interrogativ/explanatorisch). Wenn die Astrologie eine hinreichend akkurate Theorie menschlicher Persönlichkeitsstrukturen und Lebensverläufe wäre, könnte man analoge Dinge – objektiv, propositional, interrogativ – auch *faktisch* (mit dieser Theorie) verstehen. Analog dazu

⁷¹Diese Unterscheidung wird häufig getroffen (z. B. Baumberger, 2011, S. 70), teils unter anderen Bezeichnungen: Stefan Sleeuw widmet seine gesamte Dissertation faktischem bzw. symbolischem Verstehen (Sleeuw, 2023). Henk de Regt unterscheidet zwischen dem Verstehen eines Phänomens und dem Verstehen einer wissenschaftlichen Theorie (de Regt, 2017, S. 23). Und Michael Strevens grenzt ‚Verstehen-mit‘ (engl. *understanding-with*), dessen Objekt eine Theorie ist, vom Verstehen eines Phänomens mithilfe einer Erklärung ab (Strevens, 2013, S. 513). John Greco unterscheidet zusätzlich zu 4. und 5. noch ein weiteres Verstehensobjekt, nämlich die Relationen zwischen Repräsentation und Repräsentandum (Greco, 2014, S. 293). Da dieses dritte Verstehensobjekt nur relevant wird, wenn man faktisches Verstehen mithilfe symbolischen Verstehens erwirbt, liste ich es nicht als distinktes Objekt des Verstehens auf.

könnte Elif verstehen, warum das Expertensystem laut der Erklärung einer XAI-Methode eine Gastritis-Diagnose gestellt hat, ohne dass diese Erklärung wahrheitsgemäß das Verhalten des Systems beschreibt. Diese Überlegungen werden in Kapitel 7, 8 und 10 erneut wichtig, wenn ich mich mit Verstehen und Überzeugung, Faktizität bzw. Rechtfertigung auseinandersetze.

6.3 Die Objekte instrumentellen Verstehens sind komplex

In der Einleitung dieses Kapitels wurde deutlich, dass verschiedenste Dinge als Objekte instrumentellen Verstehens in Frage kommen. Da es sich jeweils um Aspekte handelt, die auf das fragliche KI-System bezogen sind, bezeichne ich diese als ‚systembezogene Aspekte‘. Manche dieser Aspekte haben einen direkten Bezug zum System. Beispielsweise können Informationen über die Zusammensetzung der Trainingsdaten, den vorgesehenen Anwendungskontext oder die Gründe, Ursachen oder Mechanismen, die zu einer bestimmten Ausgabe geführt haben, dabei helfen, die Fairness eines Systems zu beurteilen. Andere systembezogene Aspekte haben nur indirekt mit dem konkreten System zu tun. Um z. B. zu verstehen, ob ein System genau arbeitet, kann auch Hintergrundwissen zur informatischen Bedeutung von ‚Genauigkeit‘ oder zu den Anforderungen der Anwendungsdomäne erforderlich sein. Und wenn es darum geht, jemanden bei Fehlern in KI-gestützten Entscheidungen zur Rechenschaft zu ziehen, braucht man womöglich auch ein gewisses Verständnis der rechtlichen oder strukturellen Rahmenbedingungen. In solchen Fällen können auch Informationen über das soziotechnische System, in welches das KI-System eingebettet ist, relevante systembezogene Aspekte darstellen.

In den vorherigen Abschnitten habe ich aufgezeigt, dass sich Verstehensobjekte verschiedentlich systematisieren lassen. Für meine Zwecke ist es hilfreich anzunehmen, dass propositionales, interrogatives und objektuales Verstehen näherungsweise unterschiedliche Komplexitätsgrade des Verstehensobjekts abbilden (siehe Abschnitt 6.1.3). Instrumentelles Verstehen kann entsprechend sowohl propositional, interrogativ als auch objektual sein. Beispielsweise könnte die Mitarbeiterin einer Zertifizierungsstelle für KI-Systeme propositional verstehen wollen, dass ein gegebenes KI-System relevante Anforderungen z. B. in Hinblick auf Fairness, Sicherheit oder Zuverlässigkeit erfüllt. Eine Ärztin, die ein KI-Diagnose-Tool verwendet, muss vielleicht explanatorisch verstehen, warum das System eine bestimmte Diagnose gestellt hat (etwa um sicherzustellen, dass sie der Empfehlung vertrauen kann). Und eine Informatikerin möchte womöglich ein KI-System objektual verstehen, um wissenschaftliche Erkenntnisse über die Funktionsweise tiefer neuronaler Netze zu gewinnen. In all diesen Fällen versteht der fragliche Stakeholder eine Gruppe von Propositionen über relevante systembezogene Aspekte, die eine gewisse Komplexität aufweist.

Die Objekte instrumentellen Verstehens sind also wie jene von Verstehen *simpliciter* stets komplex. Nehmen wir beispielsweise an, eine Informatikerin kommt zu dem Schluss,

dass ein KI-System unfair ist (z. B. in dem Sinne, dass es die kontextuell geeignete Fairness-Metrik nicht erfüllt). Sie möchte diesen Fehler beheben. Dazu muss sie verstehen, woher das Problem rührt und was sie verändern muss, um das System fair(er) zu machen – sie braucht ein Verständnis relevanter systembezogener Aspekte. Beispielsweise könnte sie die Trainingsdaten des Systems analysieren und feststellen, dass die benachteiligte Gruppe darin unterrepräsentiert ist. Sie könnte mithilfe von XAI-Methoden herausfinden, dass das System problematische Merkmale wie das Geschlecht oder die ethnische Zugehörigkeit der Entscheidungssubjekte heranzieht. In Kombination mit ihrem Verständnis der relevanten Hintergründe – z. B. der verwendeten Lernalgorithmen und der Häufigkeit einer solchen Bias-Problematik – könnte sie verstehen, warum das System unfair ist und wie sie dagegen vorgehen kann (etwa durch die Diversifizierung und bessere Aufbereitung der Trainingsdaten). Ein solches Verständnis erfordert das Begreifen von Zusammenhängen in einer größeren Informationsmenge, keine atomistische Überzeugung über eine simple Proposition.

Doch sind Objekte instrumentellen Verstehens wirklich notwendigerweise komplex? Der folgende Fall scheint ein Beispiel für atomistisches instrumentelles Verstehen im Kontext eines Fairness-Desideratums zu sein: ‚Lanya versteht, dass das KI-System unfair ist.‘ In diesem Fall ist das Fairness-Desideratum epistemisch erfüllt, indem eine einfache (wahre) Proposition (*Das KI-System ist unfair*) geglaubt wird. Um diese Überzeugung zu erwerben, braucht Lanya kein Verständnis der Zusammenhänge, welche die Unfairness des Systems erklären oder herbeiführen. Wir können uns beispielsweise vorstellen, dass sie sich auf das Zeugnis einer vertrauenswürdigen Expertin verlässt. Warum sollte dieses atomistische, propositionale Verstehen nicht als instrumentelles Verstehen durchgehen?

Der Grund dafür sind die Adäquatheitskriterien hinsichtlich Ähnlichkeit und Nützlichkeit, die mein Explikat erfüllen muss (Abschnitt 4.3). Erstens wird die ausreichende semantische Ähnlichkeit zwischen Explikat und Explikandum gefährdet, wenn man atomistische Verstehensobjekte zulässt. Das Begreifen von Zusammenhängen in komplexen Informationsmengen ist ein zentrales Wesensmerkmal von Verstehen *simpliciter* – so zentral, dass mit Verstehen befasste Erkenntnistheoretiker:innen als atomistisch aufgefasstes propositionales Verstehen zumeist mit propositionalem Wissen gleichsetzen und als Untersuchungsgegenstand verwerfen (siehe Abschnitt 6.1). In der XAI-Debatte wird hingegen dezidiert auf Verstehen (und nicht Wissen) verwiesen (siehe Abschnitte 4.1.1 und 4.2.3), und Erklärungen sind auch außerhalb von KI-Erklärbarkeit traditionell eng mit Verstehen (und weniger mit Wissen) assoziiert (Abschnitt 3.3.2). Dies spricht dafür, Objekte instrumentellen Verstehens als komplex aufzufassen.

Wichtiger ist jedoch, zweitens, der theoretische Zweck des Explikats. Der wichtigste Zweck meines Begriffs instrumentellen Verstehens besteht darin, die Mediatorfunktion zu erklären, die Verstehen in XAI-Kontexten zugeschrieben wird. Warum kann Verstehen dabei helfen, eine Vielzahl gesellschaftlicher Desiderata in XAI-Kontexten epistemisch (und mitunter auch substanziell) zu erfüllen? Fälle wie jener von Lanya, in denen jemand lediglich eine atomistische Überzeugung über den Status eines Desideratums erwirbt,

werden dieser explanatorischen Funktion nicht gerecht. Explanatorisch ist in diesem Fall nicht das (vorgebliche) Verstehen von Lanya, sondern das Verständnis der Expertin, auf deren Zeugnis sie sich stützt. Anders ausgedrückt: Fälle, in denen ein gegebenes Desideratum durch eine atomistische Überzeugung erfüllt wird, setzen voraus, dass jemand anderes ein komplexeres Verständnis besitzt, von dem andere Personen profitieren können.

Zweifelsohne muss nicht jeder Stakeholder seine Desiderata durch instrumentelles Verstehen erfüllen – ‚instrumentelles Wissen‘ reicht oftmals aus.⁷² Diese triviale Einsicht geht aber an der eigentlichen Herausforderung von Erklärbarkeit vorbei. Natürlich kann man durch das Zeugnis von Expert:innen testimoniales Wissen über den Status eines Desideratums erwerben. Doch dieses epistemische Ziel motiviert nicht die Erklärbarkeitsforschung. Die eigentliche (und epistemologisch interessante) Schwierigkeit besteht darin, die ‚Black Box‘ so weit zu öffnen, dass verschiedene Stakeholder relevante Aspekte von KI-Systemen verstehen können. Es ist dieses Verstehen, das ich in dieser Arbeit näher beleuchte.

Für instrumentelles Verstehen ist die Unterscheidung zwischen propositionalem, interrogativem und objektuaalem Verstehen daher nachrangig. Entscheidend ist, dass Objekte instrumentellen Verstehens Propositionen über systembezogene Aspekte sind, über die ein hinreichend komplexes Überzeugungssystem ausgebildet werden kann.

Allerdings wurde in Abschnitt 6.1 deutlich, dass sich Verstehensobjekte nicht bloß als unterschiedlich komplexe Mengen von Propositionen auffassen lassen. Zwar hat sich die Unterscheidung anhand des grammatikalischen Objekts nicht als hilfreich erwiesen (Abschnitt 6.1.1). Doch habe ich in Abschnitt 6.1.2 herausgearbeitet, dass die Differenzierung anhand des propositionalen bzw. nicht-propositionalen Charakters des Verstehensobjekts auf relevante Aspekte von Verstehen hinweist: Es ist nicht offensichtlich, dass das Begreifen von Zusammenhängen, verstehensbezogene Fähigkeiten, die Verwendung nicht-propositionaler Verstehensermöglicher sowie die etwaige phänomenale Komponente mancher Fälle von Verstehen durch adäquate propositionale Einstellungen gegenüber einer hinreichend komplexen Informationsmenge abgedeckt werden.

Daher trage ich diesen Aspekten an anderer Stelle Rechnung. In Abschnitt 7.3 werde ich mehrere Möglichkeiten aufzeigen, wie der Übergang von nicht-propositionalen Repräsentationen oder Dingen in der Welt zu propositionalen Einstellungen aufgefasst und erklärt werden kann. Dass nicht-propositionale Dinge verstanden werden können, verhindert demnach nicht, dass Verstehen ausschließlich propositionale Einstellungen erfordert. Außerdem werde ich in Kapitel 9 mit verstehensbezogenen Fähigkeiten eine

⁷²Wenn Adam Carter und Emma Gordon Recht haben, ist das jedoch nicht immer der Fall. Den Autor:innen zufolge ist der Besitz von Verstehen manchmal eine relevante epistemische Norm des Behauptens (Carter & Gordon, 2011; siehe Fußnote 45). Demnach würde es für eine Ärztin nicht genügen zu wissen, dass eine KI-Diagnose zutreffend ist, wenn sie dies z. B. gegenüber Patient:innen oder Kolleg:innen behauptet. Um korrekterweise zu behaupten, dass die Diagnose stimmt, müsste sie auch verstehen, warum dem so ist. Der Grund ist, dass wir von einer Ärztin legitimerweise eine entsprechende Expertise erwarten (siehe Abschnitt 4.2.1). Analog könnte es sein, dass man von manchen Stakeholdern nicht bloß ‚instrumentelles Wissen‘, sondern instrumentelles Verstehen erwarten darf.

mutmaßlich nicht rein propositionale Komponente in meinen Verstehensbegriff integrieren – auch wenn ich es plausibel finde, dass diese Fähigkeiten letztlich aus einem adäquat verknüpften System an (propositionalen) Überzeugungen erwachsen. In diesem Zuge werde ich auch einen Vorschlag dazu skizzieren, wie das womöglich mit Verstehen assoziierte phänomenale Erleben in die Fähigkeitsbedingung von Verstehen integriert werden könnte (siehe Fußnote 108).

Mein Vorschlag fügt sich in Kenneth Boyds Unterscheidung zwischen der Inhalts- und Begreifenskomponente von Verstehen ein (Boyd, 2017). Die *Inhalte* instrumentellen Verstehens sind propositional. Diese Propositionen müssen miteinander verknüpft sowie hinreichend komplex und umfangreich sein, um als Verstehensobjekte in Frage zu kommen. Diese Perspektive ist kompatibel damit, dass die *Begreifenskomponente* instrumentellen Verstehens nicht auf propositionale Einstellungen reduziert werden kann, und dass nicht-propositionale Dinge propositional verstanden werden können.

Was ist mit faktischem und symbolischem Verstehen? Mein Vorschlag in Abschnitt 6.2 lautete, dass diese Kategorien orthogonal zu den drei anderen Objektarten verlaufen und die Repräsentationsebene des Verstehensobjekts betreffen. Ist das instrumentelle Verstehen systembezogener Aspekte also als symbolisch oder faktisch aufzufassen? Das hängt davon ab, wie man diese Kategorien auslegt. Denn in XAI-Kontexten haben wir es nicht zwangsläufig mit einer einfachen Repräsentations-Relation zu tun, wie es zum Beispiel bei einer wissenschaftlichen Theorie (Repräsentation) über ein natürliches Phänomen (Repräsentandum) der Fall ist. Bei KI-Systemen handelt es sich gewissermaßen um ‚repräsentationale Artefakte‘. Illustrieren wir das am Beispiel eines ML-basierten Systems:

Repräsentationsebenen in XAI-Kontexten: Ein *Verstehensermöglicher* repräsentiert Aspekte eines *KI-Systems*, das Aspekte der *Trainingsdaten* repräsentiert, die wiederum Aspekte der *Welt* repräsentieren.

Aufgrund dieser komplexen Repräsentationsbeziehungen, die in XAI-Kontexten zum Tragen kommen können, kann man instrumentelles Verstehen systembezogener Aspekte häufig sowohl als faktisch als auch symbolisch auffassen. Verdeutlichen wir dies an einem einfachen Beispiel, in dem die einzigen relevanten Verstehensobjekte Aspekte sind, die sich direkt auf ein KI-System beziehen. Der ersten Lesart zufolge sind KI-Systeme Artefakte und damit Dinge in der Welt. Instrumentelles Verstehen wäre demnach faktisch. Verstehensermöglicher, die selbst wiederum Gegenstand symbolischen Verstehens sein können, spielen dann die Rolle des Vehikels für den Erwerb dieses faktischen Verstehens. Beispielsweise dient eine Salienzkarte als symbolisches Vehikel, um damit faktisch zu verstehen, warum ein Bildklassifikationssystem das Foto eines Kettenkarussells als ‚Hängebrücke‘ klassifiziert hat (siehe Abbildung 3.3b).

Die zweite Lesart trägt dem Umstand Rechnung, dass KI-Systeme selbst als Repräsentationen von Dingen in der Welt aufgefasst werden können. Konventionelle KI repräsentiert diese Dinge z. B. durch Programmcode oder eine Wissensbasis. ML-basierte Systeme sind repräsentational, weil sie mit Daten trainiert werden, die Teile der Welt repräsentieren.

Beispielsweise repräsentiert ein CNN-basiertes Bildklassifikationssystem die Bilder, mit denen es trainiert wurde – und bestenfalls auch dasjenige, was diese Bilder zeigen –, in Form einer numerischen Repräsentation (siehe Abschnitt 2.1.2). Dieser Lesart zufolge ist instrumentelles Verstehen symbolisch, und wird erst faktisch, wenn man mithilfe des KI-Systems einen Aspekt der Welt versteht. Das ist unter anderem in manchen wissenschaftlichen Kontexten der Fall, wenn Wissenschaftler:innen Computersimulationen oder andere KI-Systeme verwenden, um damit natürliche Phänomene oder Strukturen zu verstehen (siehe etwa Sullivan, 2022).

Beide Lesarten haben ihre Berechtigung. Doch ist deutlich geworden, dass die genannten Oberkategorien uneindeutig sind und das jeweilige Verstehensobjekt daher explizit gemacht werden muss. Sofern nicht anders angegeben, fasse ich instrumentelles Verstehen als holistisches, faktisches Verstehen systembezogener Aspekte eines KI-Systems auf, üblicherweise vermittelt durch das symbolische Verstehen eines Verstehensermöglichers. Nun, da feststeht, worin die Objekte instrumentellen Verstehens bestehen, können wir uns der Einstellung zuwenden, die ein instrumentell verstehender Stakeholder gegenüber diesen Inhalten einnimmt.

7 Verstehen und Überzeugung

Überzeugung ist notwendig für Wissen – darin ist sich die Mehrheit der Philosoph:innen einig. Um beispielsweise zu wissen, dass ein medizinisches Expertensystem eine bestimmte Diagnose gestellt hat, muss ich notwendigerweise auch glauben, dass dem so ist. Plausiblerweise gilt dasselbe für Verstehen: Wenn ich verstehe, dass das Expertensystem eine Gastritis diagnostiziert hat, weil es eine Helicobacter-Infektion detektiert hat, dann glaube ich auch, dass das System diese Diagnose gestellt hat, weil es eine solche Infektion detektiert hat. Es wäre seltsam, zu sagen, dass ich verstehe, dass das System eine Gastritis diagnostiziert hat, weil es eine Helicobacter-Infektion detektiert hat, aber dass ich nicht glaube, dass dem so ist. Entsprechend ist es kaum überraschend, dass eine Reihe von Philosoph:innen annimmt, dass Verstehen, genau wie Wissen, Überzeugung erfordert. Autor:innen, denen zufolge Verstehen eine Form des Wissens ist oder darauf reduziert werden kann, verpflichten sich zwangsläufig auf diese Position (etwa Grimm, 2006; Riaz, 2015; Sliwa, 2015). Aber auch viele Philosoph:innen, denen zufolge Wissen nicht notwendig und/oder hinreichend für Verstehen ist, nehmen an, dass Überzeugung die in Verstehen involvierte epistemische Einstellung ist (z. B. Hills, 2016; Kvanvig, 2003; Pritchard, 2009).⁷³

Allerdings wurden im Zuge der Versuche, Verstehen von Wissen abzugrenzen, auch Beispiele vorgebracht, die zeigen sollen, dass Überzeugung – anders als es zunächst scheint – nicht notwendig für Verstehen ist. Im Anschluss an diese Beispiele wird häufig vorgeschlagen, dass *Akzeptanz* eine geeignete Alternative darstellt. Meine Position ist, dass solche Alternativen überflüssig sind, weil Überzeugung notwendig für (instrumentelles) Verstehen ist. Die eingangs formulierte Intuition ist also korrekt. Um diese Position zu verteidigen, weise ich in diesem Kapitel die (vorgeblichen) Beispiele für Verstehen ohne Überzeugung zurück.

Es lassen sich drei Kategorien von Fällen unterscheiden, die ich in den folgenden Abschnitten nacheinander diskutieren werde: Erstens wurden Beispiele vorgebracht, in denen eine Theorie bzw. Erklärung nicht geglaubt, aber trotzdem verstanden wird. Solche Fälle betrachte ich in Abschnitt 7.1. Zweitens werden Beispiele diskutiert, in denen eine teils falsche Theorie nicht geglaubt, aber verstanden wird, und außerdem genutzt wird, um damit einen Aspekt der Welt zu verstehen (Abschnitt 7.2). Schließlich führen manche Autor:innen Fälle an, in denen eine Person etwas Nicht-Propositionales (z. B. visuelle Repräsentationen) versteht, aber nicht glaubt. Diesen Beispielen widme ich mich in Abschnitt 7.3. Abschließend übertrage ich die gewonnenen Einsichten auf instrumentelles Verstehen (Abschnitt 7.4).

⁷³Ich komme in Abschnitt 11.1.3 auf das Verhältnis von Wissen und Verstehen zurück.

7.1 Verstehen einer Theorie

Die erste Kategorie von vorgeblichen Beispielen für Verstehen ohne Überzeugung umfasst Fälle, in denen eine Person eine Theorie oder Erklärung nicht glaubt, aber trotzdem versteht. Schauen wir uns also einige Fallbeispiele aus der Literatur an. Den ersten Fall übernehme ich von Daniel Wilkenfeld. Er ist angelehnt an das *Challenger*-Unglück im Jahr 1986, bei dem eine Raumfähre der NASA kurz nach dem Start in der Luft zerbrach und alle sieben Besatzungsmitglieder ums Leben kamen, weil mindestens ein Dichtungsring an einem Raketentriebwerk versagte:

Challenger:⁷⁴ „Richard is an established and skilled scientist tasked with investigating why the Challenger space shuttle exploded. His investigative instincts are excellent, and he is immediately suspicious of the O-Rings’ ability to operate at cold temperatures. In fact, the reason the Challenger exploded was because of O-Ring failure. Sadly, before he goes public he is subjected to a deliberate cover-up, casting doubt on whether the O-Rings failed in this case. Arbitrarily large amounts of evidence can be presented that there really was no O-Ring failure, to the point that Richard loses confidence in his belief that the Challenger exploded because of O-Ring failure. [...] However, Richard still has a detailed model of how O-Ring failure would cause the explosion. Moreover, when he pictures the explosion, it is often accompanied by the fleeting suspicion – which he would still not put at above 30 percent – that the O-Rings might have failed. [...] Richard is, in fact, the world’s preeminent expert on the cause of the Challenger explosion – he just fails to believe his own conclusions.“ (Wilkenfeld, 2017b, S. 321–322)

Der zweite Fall stammt von Finnur Dellsén und betrifft die Einstellung einer Physikerin zur Stringtheorie. Diese Theorie nimmt unter anderem an, dass so genannte ‚Strings‘ und ‚Brane‘ die fundamentalen Bausteine unseres Universums darstellen:

Zweifelnde Wissenschaftlerin:⁷⁵ „Carrie is a theoretical physicist in a nearby possible world (perhaps this one) in which string theory is true. Carrie has built her career around using string theory to explain various known phenomena about the

⁷⁴Malfatti beschreibt zwei ähnlich gelagerte Fälle. Im ersten Fall gelingt es dem außerordentlich erfolglosen Kriminalbeamten Klaus einmalig, den wahren Mörder und sein Tatmotiv zu identifizieren. Allerdings glaubt Klaus seine eigene Hypothese nicht, weil er um seine katastrophale Aufklärungsquote weiß (Malfatti, 2023, S. 190, 192). Im zweiten Fall führt ein irreführender Anfechtungsgrund dazu, dass eine Ärztin die wahre Erklärung für die Beschwerden eines Patienten (nämlich Eisenmangel) nach ihrer anfänglich korrekten Einschätzung wieder verwirft (ebd., S. 188).

⁷⁵Einen ähnlichen Fall beschreiben Adam Carter und Emma Gordon (die wie ich annehmen, dass Verstehen Überzeugung erfordert). Die Protagonistin ihres Beispiels ist die nihilistische Professorin Bella, die sich aus dem Wunsch nach finanzieller Sicherheit heraus dafür entschieden hat, Professorin für Evolutionsbiologie zu werden. Persönliches Interesse bringt sie für das Thema allerdings nicht auf. Obwohl sie eine Expertin für die Evolutionstheorie ist, hat sie sich daher nie dafür interessiert, ob die Theorie wahr ist, und entsprechend auch keine Überzeugung diesbezüglich ausgebildet (Carter & Gordon, 2016, S. 435).

natural world, and has become one of the world's leading contributors in the field because of her unmatched insight into the theory and its applications. Moreover, she has adopted the policy of treating string theory as given in her scientific endeavours – using it in explanations of various natural phenomena – and thus accepts string theory for explanatory purposes. However, like many other physicists, Carrie has significant methodological reservations about string theory in its current form, and therefore is not disposed to feel that string theory is even approximately true. In other words, Carrie does not believe that string theory is even approximately true.“ (Dellsén, 2017, S. 249)

Ich werde unten argumentieren, dass sich die Beispiele der ersten Kategorie, einschließlich *Challenger* und *Zweifelnde Wissenschaftlerin*, mithilfe der in Abschnitt 6.2 getroffenen Unterscheidung zwischen faktischem und symbolischem Verstehen zurückweisen lassen. Doch scheitert zumindest das *Challenger*-Beispiel meines Erachtens bereits aus einem viel simpleren Grund: Ich finde es schlichtweg unplausibel, dass Richard versteht, warum die Raumfähre explodiert ist. Im Anschluss an Michael Hannon lässt sich diese Einschätzung erklären, indem man die vorgeblich verstehende Person näher betrachtet: Wenn man Richard nach einer Erklärung für die Explosion fragte, würde er nicht das Versagen der Dichtungsringe angeben, sondern sich enthalten oder eine falsche Erklärung benennen. Wenn Richard *selbst* nicht die korrekte Ursache für das zu erklärende Ereignis identifiziert, wie sollte er dann verstehen, warum dieses Ereignis eintrat (Hannon, 2021, S. 272)? Vielmehr scheint es so zu sein, dass die wahre, von Richard verworfene Erklärung prinzipiell geeignet ist, die Challenger-Explosion zu verstehen – aber lediglich für andere Personen, die nicht von der Vertuschung betroffen sind. Richard selbst versteht aber nicht (Baumberger, 2019, S. 385). *Challenger* kann meiner Ansicht nach daher nicht zeigen, dass Überzeugung nicht notwendig für Verstehen ist.

Doch was ist mit *Zweifelnde Wissenschaftlerin*? Carrie verhält sich, anders als Richard, durchaus so, als würde sie die Stringtheorie glauben: Wenn ihre Studierenden oder Kolleg:innen sie nach der Anzahl der Dimensionen unseres Universums oder dem Wesen der Gravitation fragen, gibt Carrie darauf aller Wahrscheinlichkeit nach eine stringtheoretische Antwort. Dellsén schlägt auf Grundlage von *Zweifelnde Wissenschaftlerin* vor, dass Verstehen anstatt Überzeugung mitunter lediglich Akzeptanz im Sinne von Jonathan Cohen (1995) involviere: „[...] understanding something may merely involve treating certain propositions or theories as given in the context of explaining something, as opposed to being disposed to feel that the propositions or theories are true“ (Dellsén, 2017, S. 248). (Da sich viele Philosoph:innen in der Verstehensdebatte auf Cohens Akzeptanzbegriff berufen, diskutiere ich diesen Vorschlag unten ausführlicher.) Dellsén stellt außerdem klar, dass Verstehen nicht nur mit *Enthaltung* kompatibel sei, wie sie von Carrie an den Tag gelegt wird. Auch, wenn Carrie glauben würde, dass die Stringtheorie falsch ist, würde sie die Theorie verstehen – solange sie die Theorie akzeptiert (ebd., S. 250).

Wenn Dellsén Recht hat, könnte auch instrumentelles Verstehen mitunter lediglich

Akzeptanz erfordern, da die Nutzer:innen von Verstehensmöglichkeiten diese nicht zwangsläufig für wahr halten. Hier ist ein analoger Fall aus dem XAI-Kontext:

XAI-Skepsis: Cynthia ist eine renommierte XAI-Expertin, die sich außergewöhnlich gut mit informatischen Grundlagen und Anwendungen erklärbarer KI auskennt. Sie forscht schon seit vielen Jahren zu Erklärbarkeit und hat mehrere einflussreiche XAI-Ansätze entwickelt, darunter auch eine lokale Post-hoc-Methode namens *Insight*. Als Entwicklerin der Methode ist Cynthia die führende Expertin für *Insight* und hat umfassende Einsicht in deren Funktionsweise und die generierten Erklärungen. Allerdings hat Cynthia, wie viele andere XAI-Forscher:innen, starke Vorbehalte gegenüber Post-hoc-Ansätzen entwickelt und ist zu dem Schluss gekommen, dass diese keineswegs getreulich wiedergeben, was tatsächlich zu einer bestimmten KI-Ausgabe geführt hat. Daher hält Cynthia auch die von *Insight* generierten Erklärungen nicht für wahr. Dennoch nutzt Cynthia *Insight*, um die Ausgaben tiefer neuronaler Netze zu erklären, und verwendet diese Erklärungen in ihrer Forschung (vielleicht aus einem Mangel an Alternativen oder weil sie ihrer Karriere nicht schaden will). Was Cynthia nicht ahnt: *Insight* liefert in der Tat ausschließlich korrekte Erklärungen.

So wie Carrie aus *Zweifelhafte Wissenschaftlerin* hält auch Cynthia aus *XAI-Skepsis* die Erklärungen, auf die sie sich in ihrer Forschung stützt, nicht für wahr, obwohl sie diesbezüglich umfassende Expertise besitzt (und die Erklärungen tatsächlich wahr sind). Daher würde Dellsén vermutlich auch Cynthia Verstehen in Abwesenheit von Überzeugung zuschreiben. Ich denke ebenfalls, dass Carrie und Cynthia Verstehen besitzen. Ich denke allerdings nicht, dass die Fälle zeigen können, dass Verstehen keine Überzeugung erfordert. Im Rahmen meiner Diskussion der dritten Art von Beispielen werde ich die von Dellsén und anderen als Alternative zu Überzeugung vorgeschlagene Akzeptanz kritisieren. An dieser Stelle möchte ich lediglich aufzeigen, dass Carries und Cynthias Verstehen durchaus Überzeugungen erfordert, dass die Berufung auf Akzeptanz also unnötig ist. Meinem Vorschlag zufolge versteht Carrie die Stringtheorie, aber sie versteht (z. B.) nicht den fundamentalen Aufbau unseres Universums; und Cynthia versteht die von *Insight* generierten Erklärungen, aber nicht die Ursachen, Gründe (o. ä.) für die damit erklärten Systemausgaben. Diese Differenzierung beruht auf der in Abschnitt 6.2 getroffenen Unterscheidung zwischen faktischem und symbolischem Verstehen.

Faktisches Verstehen hat als Objekt ein Faktum, Ereignis oder Ding – zum Beispiel den fundamentalen Aufbau unseres Universums oder die tatsächlichen Vorgänge in einem KI-System. Das Objekt symbolischen Verstehens ist hingegen eine symbolische Repräsentation, z. B. linguistische Erklärungen, Theorien, Abbildungen oder Diagramme – beispielsweise die Stringtheorie oder eine XAI-Erklärung. Mithilfe dieser Unterscheidung lässt sich leicht erklären, warum wir Carrie und Cynthia intuitiv Verstehen zugestehen: Carrie besitzt (objektuales) symbolisches Verstehen der Stringtheorie. Sie kennt die Theorie also sehr genau und begreift, wie ihre einzelnen Elemente zusammenhängen. Sie kann von der

Stringtheorie angenommene Phänomene (z. B. die Gravitation) mithilfe der Theorie erklären und kompetent verschiedene Fragen zur Stringtheorie beantworten. Der Grund dafür ist, dass sie eine Reihe von Überzeugungen hinsichtlich der Stringtheorie besitzt – nämlich solche, die (implizit) den Zusatz ‚Laut der Stringtheorie‘ beinhalten. Beispielsweise glaubt Carrie, dass *laut der Stringtheorie* unser Universum aus Strings und Branen zusammengesetzt ist. Allerdings hält sie die Theorie nicht für wahr. Daher glaubt sie nicht, dass unser Universum *de facto* aus Strings und Branen zusammengesetzt ist. Und wenn Carrie das nicht glaubt bzw. wenn sie *die Stringtheorie nicht einmal für annähernd wahr hält*, ist es meines Erachtens unplausibel, dass sie die Zusammensetzung unseres Universums (mithilfe der Stringtheorie) versteht. Was Carrie also fehlt, ist (objektuales) faktisches Verstehen der Beschaffenheit unserer physischen Welt.⁷⁶ Analoges gilt für Cynthia aus *XAI-Skepsis*: Sie versteht die von *Insight* generierten Erklärungen (symbolisch) sehr genau und besitzt diverse ‚Laut Erklärung E' ‘-Überzeugungen. Doch wenn Cynthia diese Erklärungen nicht für eine korrekte Beschreibung der Systemvorgänge hält, die zu den fraglichen Ausgaben geführt haben, versteht sie damit nicht, was (faktisch) im System vor sich geht – auch, wenn die Erklärungen eigentlich wahr sind.

Die Differenzierung zwischen symbolischem und faktischem Verstehen ist in meinen Augen attraktiv, weil sie die Intuition erklären kann, dass Carrie und Cynthia jeweils einen hohen Verstehensgrad besitzen. Zudem lässt sich damit auch das *Challenger*-Beispiel erklären: Richard hat ein tiefes (symbolisches) Verständnis des Zusammenhangs zwischen dem Versagen der Dichtungsringe an der Raumfähre und ihrem Auseinanderbrechen. Was ihm fehlt, ist die Überzeugung, dass die Erklärung (hinreichend) wahr ist. Richard ist also genau wie Carrie und Cynthia auf der symbolischen Verstehensebene stehen geblieben. Er benutzt die Theorie bzw. Erklärung nicht, um damit einen Aspekt der Welt zu verstehen.

Für die Frage nach Verstehen und Überzeugung bedeutet das Folgendes: Eine Person, die z. B. die Erklärung ‚ p weil q ‘ mithilfe von Theorie T symbolisch versteht, besitzt die Überzeugung, dass p weil q laut T . Sie glaubt also nicht zwangsläufig, dass q tatsächlich die Ursache für p ist. Symbolisches Verstehen erfordert lediglich, dass die von einer Theorie, Erklärung o. ä. postulierten Zusammenhänge verstanden und *im Kontext* der Theorie, Erklärung o. ä. für wahr gehalten werden. In verallgemeinerter Form erhalten wir folgende

⁷⁶Ich habe dieses Argument selbst entwickelt, aber später bemerkt, dass Seungbae Park einen ähnlichen Punkt macht, um Dellsens Carrie-Fallbeispiel zu kritisieren: Park differenziert zwischen dem Verstehen der Welt und dem Verstehen einer Theorie und spricht Carrie ersteres ab (Park, 2017, S. 385). Ich interpretiere Park so, dass er darüber hinaus drei weitere Gegenargumente formuliert. (1) Um zu verstehen, müsse Carrie behaupten, dass p . Wenn sie nicht glaubt, dass p , lande man in einem Moor'schen Paradox (ebd., S. 383). (2) Wenn Carrie das Explanandum einer stringtheoretischen Erklärung glaube, nicht aber das Explanans, folge sie einem problematischen Doppelstandard (ebd., S. 383–384). (3) Akzeptanz entkoppele Verstehen von den Gegebenheiten der Welt (ebd., S. 384). Akzeptanz könne daher keine Handlungen anleiten (ebd., S. 385) und weder wissenschaftlichen Fortschritt (ebd., S. 387) noch das Phänomen von ‚Missverstehen‘ (ebd., S. 386) erklären. Ich komme in Abschnitt 10.1 erneut auf Parks drittes Argument zurück. Auch Malfatti, die gegen die Notwendigkeit von Überzeugung argumentiert, gesteht ein, dass das ‚Laut X' -Schema manche Fälle von Verstehen ohne Überzeugung entkräften kann (Malfatti, 2023, S. 193).

notwendige Bedingung für symbolisches Verstehen:

Symbolisches Verstehen: Ein Subjekt S versteht eine symbolische Repräsentation R mit propositionalem Gehalt p nur dann, wenn S glaubt, dass p laut R .⁷⁷

Durch die Unterscheidung zwischen faktischem und symbolischem Verstehen lassen sich auch Fälle einfangen, in denen eine bekanntermaßen falsche Theorie (z. B. die Phlogistontheorie oder Astrologie) verstanden wird (Malfatti, 2023, S. 53–54), oder eine fiktionale Geschichte (Carter & Gordon, 2016, S. 427). Symbolisches Verstehen ist in diesen Fällen möglich, auch ohne die Theorien bzw. Geschichten für wahr zu halten und damit faktisches Verstehen über einen Aspekt der Welt zu erwerben. Zusammengefasst: Die Differenzierung zwischen symbolischem und faktischem Verstehen zeigt, dass die erste Gruppe von Fällen keine Beispiele für (faktisches) Verstehen ohne Überzeugung liefert, sondern lediglich für (symbolisches) Verstehen mit Überzeugung.

7.2 Verstehen mithilfe einer Theorie

Die zweite Kategorie von Fällen beschreibt Situationen, in denen eine Person, normalerweise in einem wissenschaftlichen Kontext, mithilfe einer streng genommen falschen, z. B. stark idealisierten oder abstrahierten, Theorie einen Aspekt der Welt versteht. (Wir werden unten sehen, dass analoge Fälle auch in XAI-Kontexten auftreten können.) Da die Person um die Abweichung zwischen Theorie und Wirklichkeit weiß, hält sie die Theorie nicht für wahr. Anstatt auf mehr oder weniger artifizielle philosophische Fallbeispiele stützt sich dieser Argumentationsstrang auf empirische Tatsachen. Denn wissenschaftliche Theorien und Modelle kommen selten ohne Idealisierungen oder sonstige Falschheiten aus (de Regt, 2015; Potochnik, 2015).

Im Kontext wissenschaftlicher Modelle oder Theorien handelt es sich bei Idealisierungen, Abstraktionen, Fiktionen, Approximationen oder anderen Falschheiten um zielgerichtete Abweichungen von der Wirklichkeit, d. h. um Annahmen, von denen bekannt ist, dass sie falsch sind. Eines der Standardbeispiele ist das Ideale-Gas-Modell (siehe etwa Elgin, 2017, S. 15; Mizrahi, 2012):

Ideales-Gas-Modell: Das Ideale-Gas-Modell modelliert den Zusammenhang zwischen Druck, Temperatur und Raumvolumen bezüglich eines idealen, also in dieser Form nicht existierenden Gases. Das ideale Gas besteht aus Molekülen ohne räumliche Ausdehnung, die keinerlei gegenseitige Anziehungskraft ausüben. Reale Gasmole-

⁷⁷Die Gradierbarkeit von Verstehen (siehe Kapitel 5) wird in dieser Bedingung nicht berücksichtigt, doch lässt sie sich leicht daran anpassen. Um beispielsweise Grade der Vollständigkeit und Faktizität abzubilden, könnte man den propositionalen Gehalt der Repräsentation feiner aufschlüsseln ($p_1, p_2, \dots, p_n$) und zwischen den (wahren bzw. falschen) Überzeugungen, dass p_1 laut R , dass p_2 laut R , ..., dass p_n laut R differenzieren. Um den Grad der Annäherung an die Wahrheit abzubilden, müsste man relevante Approximationen an p unterscheiden.

leküle besitzen hingegen eine räumliche Ausdehnung und ziehen sich gegenseitig an bzw. stoßen sich ab.

Physiker:innen wissen, dass zentrale Annahmen des Modells falsch sind. Dennoch wird es regelmäßig eingesetzt, um damit reale thermodynamische Prozesse zu beschreiben. Gleiches gilt für andere idealisierte Modelle, die in verschiedenen wissenschaftlichen Kontexten zum Einsatz kommen – beispielsweise werden menschliche Akteur:innen in wirtschaftswissenschaftlichen Modellen häufig als vollkommen rational aufgefasst, und physikalische Theorien lassen Reibungskräfte oft unberücksichtigt. Wissenschaftler:innen verschiedener Disziplinen entwickeln und verwenden also teils stark idealisierte oder anderweitig verfälschende Theorien und Modelle, obwohl sie wissen, dass sie falsche Annahmen enthalten. Doch wenn man weiß, dass etwas falsch ist, glaubt man es nicht: Überzeugung ist eine Einstellung, die Dinge betrifft, die man für wahr hält. Trotzdem verstehen Wissenschaftler:innen mithilfe solch falscher Theorien eine Menge über unsere Welt (Baumberger, 2019, S. 384). Analoge Fälle können plausiblerweise auch in XAI-Kontexten auftreten:

Approximation: Aysan möchte ihre Urlaubsfotos thematisch ordnen und verwendet ein CNN, um ihre Bilder zu klassifizieren. Dabei fällt ihr auf, dass das Foto eines Kettenkarussells vom System fälschlicherweise als ‚Hängebrücke‘ klassifiziert wurde. Sie generiert mithilfe von LIME eine Erklärung für die falsche Klassifikation (siehe Abbildung 3.3). Aysan ist bekannt, dass LIME keine perfekte Erklärung der Klassifikation liefert, sondern die eigentliche Funktion des neuronalen Netzwerks lokal durch eine lineare Funktion approximiert.

Will Fleisher argumentiert, dass Post-hoc-Methoden wie LIME analog zu idealisierten wissenschaftlichen Modellen funktionieren (Fleisher, 2022). Er diskutiert drei Eigenschaften von Idealisierungen, die sie nützlich für den Erwerb wissenschaftlichen Verstehens machen und die in analoger Weise auch auf Post-hoc-Ansätze zutreffen: Vereinfachung, Kennzeichnung (engl. *flagging*) und Fokussierung auf bestimmte kausale Muster. Wie auch das Ideale-Gas-Modell *vereinfacht* LIME die nicht-lineare Funktion von Aysans CNN, indem sie einen Teil davon linear approximiert. Und so wie das Ideale-Gas-Modell bestimmte Eigenschaften realer Gase (z. B. die Interaktion von Gasmolekülen) als irrelevant *kennzeichnet*, blendet Aysans LIME-Erklärung jene Bildbereiche aus, die nur einen geringen Einfluss auf die Klassifikation hatten. Schließlich rückt das Ideale-Gas-Modell bestimmte kausale Prozesse in den *Fokus* (etwa den Zusammenhang von Druck, Volumen und Temperatur) – analog dazu fokussiert die LIME-Erklärung auf bestimmte (quasi-)kausale Muster: Als lokale Methode beschränkt sich LIME auf Bildmerkmale, welche die Klassifikation einer kleinen Menge an Eingaben beeinflussen, und hebt lediglich eine Auswahl der wichtigsten Bildmerkmale hervor (ebd., Kap. 7). Post-hoc-Methoden gleichen also in relevanten Hinsichten wissenschaftlichen Idealisierungen (mehr dazu in Kapitel 12).

Wenn Beispiele wie das Ideale-Gas-Modell zeigen können, dass Verstehen mitunter ohne Überzeugung auskommt, betrifft dies auch manche Fälle von (instrumentellem) Verstehen

in XAI-Kontexten. Beispielsweise würde Aysan aus *Approximation* demnach die (falsche) Klassifikation ihres Urlaubsfotos verstehen, obwohl sie die LIME-Erklärung nicht glaubt.

Doch auch die zweite Kategorie von vorgeblichen Gegenbeispielen lässt sich zurückweisen. Anders als bei der ersten Kategorie genügt es aber nicht, sich auf die symbolisch/faktisch-Unterscheidung zu berufen und Wissenschaftler:innen jegliches faktische Verstehen abzusprechen. Offensichtlich besitzen oder erwerben Wissenschaftler:innen mitunter faktisches Verstehen, indem sie falsche Theorien, die sie auch für falsch halten, auf Aspekte der Welt anwenden. Gleichermäßen ist es plausibel, dass Personen wie Aysan mit Post-hoc-Erklärungen bis zu einem gewissen Grad verstehen können, warum KI-Systeme bestimmte Ausgaben generieren. In Bezug auf Carrie aus *Zweifelnde Wissenschaftlerin*, Richard aus *Challenger* und Cynthia aus *XAI-Skepsis* habe ich argumentiert, dass allen dreien faktisches Verstehen abgeht, weil sie nicht an die Wahrheit der jeweiligen Theorien bzw. Erklärungen glauben. Wie kommt es, dass Wissenschaftler:innen und Aysan, anders als die zuvor diskutierten Protagonist:innen, faktisches Verstehen besitzen, obwohl sie – so scheint es – in einer ähnlichen epistemischen Situation sind?

Hier kommen die Alternativen zu Überzeugung ins Spiel, die einige Philosoph:innen vorgeschlagen haben. Das Argument lautet in etwa so: Wissenschaftler:innen sind aufgrund ihrer überdurchschnittlichen Expertise in ihrem Fachbereich paradigmatische Verstehenssubjekte. Da Wissenschaftler:innen Theorien verwenden, die sie für falsch halten, kann Überzeugung nicht die (einzige) für Verstehen charakteristische epistemische Einstellung sein. Benötigt wird, wenigstens in solchen Fällen, etwas anderes – zum Beispiel Akzeptanz (Baumberger, 2019, Kap. 4.4; Elgin, 2021, S. 1580; Malfatti, 2023, S. 193–194; Potochnik, 2015, S. 75–76) oder Verpflichtung (engl. *commitment*; Baumberger & Brun, 2017). Beispielsweise begründen Christoph Baumberger und Georg Brun ihre Verpflichtungsbedingung für objektuales Verstehen folgendermaßen: „[...] since one can understand a subject matter by means of a highly idealized theory one knows to be false, the commitment condition does not require belief in the sense of taking it to be true“ (ebd., S. 169).

Wie zuvor angedeutet, berufen sich viele Gegner:innen einer Überzeugungsbedingung für Verstehen (z. B. Baumberger, 2019, S. 384; Dellsén, 2017, S. 248; Dellsén, 2018, Kap. 4.2; Elgin, 2017, S. 18–23; Malfatti, 2023, Fn. 6 auf S. 21, S. 193–194) auf Jonathan Cohen, der Überzeugung und Akzeptanz folgendermaßen voneinander abgrenzt:

[...] belief that *p* is a disposition, when one is attending to issues raised, or items referred to, by the proposition that *p*, normally to feel it true that *p* and false that *not-p*, whether or not one is willing to act, speak, or reason accordingly. But to accept the proposition or rule of inference that *p* is to treat it as given that *p*. More precisely, to accept that *p* is to have or adopt a policy of deeming, positing, or postulating that *p*—i.e. of including that proposition or rule among one’s premisses for deciding what to do or think in a particular context, whether or not one feels it to be true that *p*. (Cohen, 1995, S. 4)

Die Idee ist also, dass Wissenschaftler:innen eine falsche Theorie nicht glauben, sondern

lediglich akzeptieren, also für bestimmte epistemische oder praktische Aktivitäten bzw. Zwecke als gesetzt annehmen. Anders als Überzeugung ist Akzeptanz der willentlichen Kontrolle des epistemischen Subjekts unterworfen (ebd., §4; Schwitzgebel, 2024, Kap. 2.4). Ich kann mich zwar nicht bewusst dazu entscheiden, eine Theorie zu glauben, aber es obliegt mir, ob ich sie akzeptieren möchte. Weiterhin ist Akzeptanz stärker als Überzeugung darauf ausgerichtet, Handlungen anzuleiten (Elgin, 2017, S. 19; Schwitzgebel, 2024, Kap. 2.4). Wissenschaftler:innen akzeptieren Theorien, um Experimente zu entwickeln oder neue Hypothesen zu überprüfen. Sie akzeptieren also, im wahrsten Sinne des Wortes, bestimmte Annahmen als *Arbeitshypothesen* – um damit arbeiten zu können (Elgin, 2017, S. 20). Analog dazu könnte Aysan aus *Approximation* die LIME-Erklärung aus pragmatischen Gründen akzeptieren – etwa, weil sie für die korrekte Sortierung ihrer Fotosammlung nützlich ist. Aysan würde die Erklärung aber nicht für wahr halten, da sie um die Abweichung zwischen dem CNN und der Erklärung weiß.

Der Begriff der Akzeptanz ist zweifellos hilfreich, um manche Einstellungen gegenüber Theorien, Erklärungen oder Modellen zu charakterisieren. Gerade in der Forschung, wo viele Fragen noch unbeantwortet sind und daher auch mit unbestätigten Hypothesen gearbeitet wird, ist Akzeptanz als willentliche, pragmatische Entscheidung für eine bestimmte Annahme ein plausibler Kandidat für relevante epistemische Einstellungen von Wissenschaftler:innen. Ich bin hingegen nicht überzeugt davon, dass diese Einstellung auch charakteristisch für Verstehen ist. Betrachten wir dazu noch einmal den Unterschied zwischen symbolischem und faktischem Verstehen.

Meine Diskussion der ersten Kategorie von vorgeblichen Beispielen für Verstehen ohne Überzeugung hat ergeben, dass symbolisches Verstehen einer Theorie kompatibel mit Überzeugung ist. Auch Personen, die eine Theorie (o. ä.) lediglich akzeptieren, also nicht für wahr halten, können diese Theorie symbolisch verstehen und wahre ‚Laut Modell/Theorie $T \dots$ ‘-Überzeugungen besitzen. Eine z. B. empirisch falsche Annahme ‚ p ‘ kann zu einer wahren Aussage werden, indem man daraus ‚Laut Theorie T , p ‘ macht (siehe auch Brogaard, 2005; van Riel, 2015). Gleichermäßen kann eine Wissenschaftlerin ‚ p ‘ für falsch halten, aber ‚Laut Theorie T , p ‘ glauben. Der Zusatz ‚laut Theorie T ‘ ermöglicht, wahre Überzeugungen auf Grundlage falscher Theorien zu erwerben (Malfatti, 2023, S. 54). Für symbolisches Verstehen braucht es also keine Berufung auf Akzeptanz.⁷⁸

Gehen wir nun über zu faktischem Verstehen mithilfe einer Theorie. Offensichtlich übertragen Wissenschaftler:innen ihre Theorien auch auf deren Gegenstand, verstehen also einen Aspekt der Welt mithilfe ihrer Theorien. Nehmen wir also an, dass eine Wissenschaftlerin auf Grundlage von Theorie T eine Erklärung E akzeptiert, derzufolge p weil q . Die Wissenschaftlerin hält es also nicht für wahr, dass p weil q , sondern stützt sich lediglich (z. B. aus pragmatischen Gründen) auf diese Annahme, wenn sie über T oder

⁷⁸Natürlich ist es möglich, dass eine Person, die eine Theorie symbolisch versteht, diese auch akzeptiert. Akzeptanz ist hingegen keinesfalls *notwendig* für symbolisches Verstehen. Schließlich kann auch eine vehemente Astrologie-Gegnerin ein hochgradiges symbolisches Verständnis astrologischer Theorien besitzen, ohne diese Theorien in irgendeinem Zusammenhang auch nur im Geringsten zu akzeptieren.

E nachdenkt oder dazu forscht. Versteht diese Wissenschaftlerin faktisch, warum *p*? Ich denke nicht, dass dem so ist. Wie bereits bei *Challenger* finde ich es äußerst kontraintuitiv, anzunehmen, dass diese Wissenschaftlerin faktisch versteht, dass *p* weil *q*, wenn sie es nicht (annähernd) für wahr hält, dass *p* weil *q*.

Doch wie kann es sein, dass Wissenschaftler:innen auf Grundlage von Theorien, die sie nicht für wahr halten, Überzeugungen ausbilden? Die Antwort lautet, dass kompetente Wissenschaftler:innen idealisierte oder anderweitig verfälschende Theorien weder für gänzlich falsch halten, noch für bare Münze nehmen. Vielmehr werden bestimmte Falschheiten gezielt und wissentlich in Modelle oder Theorien integriert, weil sie verschiedene epistemische und/oder pragmatische Vorteile mit sich bringen (Fleisher, 2022; Lawler, 2021).

Wissenschaftler:innen wissen also normalerweise, wie sie ein streng genommen falsches Modell interpretieren müssen, um daraus Schlussfolgerungen über reale Phänomene in der Welt abzuleiten. Eine Physikerin, die das Ideale-Gas-Modell benutzt, glaubt natürlich nicht, dass reale Gasmoleküle keinen Raum einnehmen. Sie zieht mithilfe des Modells Schlüsse bezüglich thermodynamischer Prozesse, die vom Modell hinreichend akkurat abgebildet werden. Das Ideale-Gas-Modell ist nämlich keine haltlose Fiktion, sondern es beschreibt wahrheitsgemäß das Verhalten von realen Gasen unter genau spezifizierten Bedingungen (Mizrahi, 2012). Beispielsweise hat die Interaktion von Gasmolekülen bei geringer Dichte und hoher Temperatur keinen Einfluss auf das Zusammenspiel von Druck, Temperatur und Volumen, sodass das Ideale-Gas-Modell für solche Umstände korrekte Ergebnisse liefert. Solche Schlüsse hält die Wissenschaftlerin, sofern sie eine kompetente Nutzerin des Modells ist, für wahr – nicht nur für wahr *laut dem Idealen-Gas-Modell*.⁷⁹

Für Verstehen und Überzeugungen bedeutet das Folgendes: Wissenschaftler:innen besitzen üblicherweise symbolisches Verstehen von Modellen oder Theorien (in Form von ‚Laut Modell/Theorie *T*...‘-Überzeugungen). Um damit faktisches Verstehen realer Phänomene zu erwerben, müssen sie diese Repräsentationen angemessen interpretieren. Sie müssen erkennen, welche Elemente veridisch sind und inwiefern andere Elemente von der Wahrheit abweichen. Dadurch können sie (wahre) Überzeugungen über das Phänomen erwerben, das vom Modell bzw. der Theorie repräsentiert wird.

Analoges gilt für XAI-Kontexte, in denen bekanntermaßen idealisierte Erklärungen

⁷⁹Mein Argument geht davon aus, dass Wissenschaftler:innen grundsätzlich wahre Beschreibungen der Welt erlangen können. Die hier vertretene Position sympathisiert also mit dem wissenschaftlichen Realismus. Wissenschaftliche Realist:innen nehmen an, dass die Wissenschaft danach strebt, die Welt bzw. natürliche Phänomene durch Theorien, Modelle oder Erklärungen wahrheitsgemäß zu beschreiben. Antirealist:innen halten dieses Vorhaben in Hinblick auf Unbeobachtbares (z. B. Atome, Quarks) für aussichtslos. Sie nehmen stattdessen an, dass das Ziel von Wissenschaft darin besteht, empirisch adäquate Beschreibungen der Welt zu erhalten, die akkurate Vorhersagen über beobachtbare Phänomene erlauben. Antirealist:innen könnten daher einwenden, dass Wissenschaftler:innen zumindest Theorien über Unbeobachtbares nicht für wahr halten (sollten), sondern lediglich für empirisch adäquat. Wenn man meiner Argumentation folgt, könnten antirealistische Wissenschaftler:innen daher vieles lediglich symbolisch, nicht aber faktisch verstehen. Doch dies scheint das korrekte Ergebnis zu sein: Wenn wissenschaftliche Theorien nicht einmal annähernd wahr sind, wie sollen sie dann genutzt werden, um die Welt faktisch zu verstehen?

verwendet werden. Aysan aus *Approximation* könnte mithilfe der LIME-Erklärung beispielsweise die Überzeugung erwerben, dass die hervorgehobenen Superpixel (in etwa) den drei ausschlaggebendsten Bildmerkmalen entsprechen (ohne zu glauben, dass die anderen Bildmerkmale keinerlei Einfluss hatten). Oder sie könnte glauben, dass die Erklärung die Klassifikationskriterien des CNN für eine eng begrenzte Menge an Eingabedaten darstellt (anstatt zu glauben, dass die hervorgehobenen Merkmale relevant für alle Fotos der Bildklasse ‚Hängebrücke‘ sind). Natürlich setzt das voraus, dass Aysan eine kompetente Interpretin der LIME-Erklärung ist. Wenn sie die Erklärung falsch interpretiert und in der Folge falsche Überzeugungen ausbildet, dann scheitert ihr Verständnis womöglich an der unzureichenden Faktizität (mehr dazu in Kapitel 8) – aber nicht an fehlenden Überzeugungen. Und wenn Aysan die Erklärung für eine haltlose Fiktion hält, dann verbleibt sie auf der symbolischen Verstehensebene (siehe Abschnitt 7.1) – und besitzt zugehörige ‚Laut Erklärung E' ‘-Überzeugungen. Auch die Beispiele der zweiten Kategorie sind daher keine Fälle von Verstehen ohne Überzeugung.

7.3 Verstehen nicht-propositionaler Repräsentationen

Die dritte und letzte Kategorie umfasst Beispiele, in denen nicht-propositionale Strukturen, Objekte oder Repräsentationen verstanden werden. Beispielsweise nimmt Christoph Baumberger an, dass manche Theorien nicht-propositionale Repräsentationen, wie Diagramme oder Graphen, beinhalten oder sogar größtenteils daraus bestehen (Baumberger, 2019, S. 374; vgl. auch Riggs, 2003, S. 218). Und es erscheint unzweifelhaft, dass Theorien verstanden werden können. Linda Zagzebski nennt diverse weitere Beispiele nicht-propositionaler Verstehensobjekte: Automotoren, Musikstücke, Kunstwerke, den Charakter eines Menschen, den Grundriss einer Stadt oder die Wirklichkeit an sich (Zagzebski, 2001, S. 242). Vor diesem Hintergrund charakterisiert sie Verstehen, wie in Abschnitt 6.1.2 erwähnt, als „the state of comprehension of nonpropositional structures of reality“ (ebd.; vgl. auch Grimm, 2014; Zagzebski, 2019).

Einige Autor:innen sehen darin einen Einwand gegen die Position, dass Verstehen stets Überzeugung involviert. Demnach wird Überzeugung, also eine propositionale Einstellung, dem Verstehen nicht-propositionaler Verstehensobjekte nicht gerecht. Eine propositionale Einstellung ist der mentale Zustand, den ein Subjekt besitzt, das eine bestimmte Einstellung, Meinung oder Haltung zu einer Proposition hat – zum Beispiel Ablehnung, Hoffnung oder Befürchtung. Überzeugung bezeichnet üblicherweise die propositionale Einstellung, eine Proposition für wahr zu halten (Schwitzgebel, 2024, o. S.). Wenn ich z. B. glaube, dass das Expertensystem eine Gastritis diagnostiziert hat, dann halte ich es für wahr, dass das System eine Gastritis diagnostiziert hat.

Das Problem ist nun Folgendes: Es ist unklar, wie propositionale Einstellungen, und damit auch Überzeugungen, bezüglich nicht-propositionaler Repräsentationen, Dinge oder Strukturen gebildet werden können. Wenn (propositionale) Überzeugung notwendig für

Verstehen ist, wie kann man dann Dinge verstehen, die keine propositionalen Inhalte haben? Nennen wir diese Problematik das *Problem der Propositionalität*. Unter anderem auf dieser Grundlage konkludiert beispielsweise Christoph Baumberger, dass die in Verstehen involvierte epistemische Einstellung Akzeptanz sei: „[...] belief is a propositional attitude, but the objects of acceptance include non-propositional contents and all kinds of theories and theory-like representations“ (Baumberger, 2019, S. 384).⁸⁰

Nicht-propositionale Verstehensobjekte, insbesondere visuelle Repräsentationen, sind äußerst relevant für das Verstehen von KI-Systemen bzw. systembezogenen Aspekten. Visualisierungen (z. B. Graphen, Bilder, Diagramme) sind gängige Verstehensmöglichkeiten in XAI-Kontexten und damit wichtige Vehikel für den Erwerb instrumentellen Verstehens. Ein typisches Beispiel sind die Salienzkarten, die häufig zur Erklärung der Ausgaben von Bildklassifikationssystemen herangezogen werden (siehe Abbildungen 3.2b und 3.3b). Sollte sich herausstellen, dass das Verstehen nicht-propositionaler Repräsentationen Akzeptanz statt Überzeugung erfordert, beträfe dies voraussichtlich auch instrumentelles Verstehen.

Allerdings werde ich nun argumentieren, dass auch die Existenz nicht-propositionaler Verstehensobjekte mit einer Überzeugungsbedingung für Verstehen kompatibel ist. Ich beschränke mich auf das Beispiel visueller Repräsentationen, da es sich angesichts der weiten Verbreitung visueller Erklärungen dabei um die in XAI-Kontexten relevanten nicht-propositionalen Verstehensobjekte handelt. In Abschnitt 7.3.1 präsentiere ich drei Lösungsstrategien für das Problem der Propositionalität. Schließlich schlage ich in Abschnitt 7.3.2 vor, dass die viel bedeutsamere Problematik bei der Verwendung visueller Repräsentationen in XAI-Kontexten im *Problem der Extraktion* besteht.

7.3.1 Lösungsstrategien für das Problem der Propositionalität

Das Argument hinter dem Problem der Propositionalität lautet in etwa so: Einerseits sind visuelle Repräsentationen mögliche und zudem gängige Verstehensobjekte, andererseits haben solche Repräsentationen nicht-propositionale Inhalte. Da Überzeugung als propositionale Einstellung ausschließlich Propositionen zum Inhalt hat, kann die in das Verstehen visueller Repräsentationen involvierte epistemische Einstellung nicht Überzeugung sein. Es braucht eine Alternative – z. B. Akzeptanz.

Zunächst einmal ist festzuhalten, dass zumindest Cohens Akzeptanzbegriff, auf den sich die Mehrheit der Gegner:innen einer notwendigen Überzeugungsbedingung stützt, nicht auf nicht-propositionale Inhalte ausgelegt ist. Hier ist noch einmal ein Auszug seiner Charakterisierung:

[...] to accept the proposition or rule of inference that p is to treat it as given that p .
More precisely, to accept that p is to have or adopt a policy of deeming, positing, or

⁸⁰ Auch Malfatti und Hubert & Malfatti scheinen anzudeuten, dass sie dies als einen Vorteil von Verpflichtung bzw. Akzeptanz gegenüber Überzeugung ansehen. Die Formulierungen sind aber uneindeutig (Hubert & Malfatti, 2023, Fn. 11 auf S. 596; Malfatti, 2023, S. 193).

postulating that p —i.e. of including that proposition or rule among one's premisses for deciding what to do or think in a particular context, whether or not one feels it to be true that p . (Cohen, 1995, S. 4)

Cohen bestimmt Akzeptanz also, genau wie Überzeugung, als Einstellung gegenüber einer Proposition p . Und auch Prämissen, zu denen die akzeptierten Propositionen bzw. Schlussregeln Cohen zufolge werden, sind paradigmatischerweise propositional.⁸¹ Laut Cohen unterscheidet sich Akzeptanz von Überzeugung also nicht etwa dadurch, dass erstere auch nicht-propositionale Inhalte haben kann. Der Unterschied ist, dass Akzeptanz für Cohen – anders als Überzeugung – nicht erfordert, dass man die fragliche Proposition für wahr hält. Ohne zusätzliche Erläuterung bleibt daher offen, wieso Cohens Akzeptanz besser geeignet sein sollte, um Einstellungen gegenüber nicht-propositionalen Repräsentationen zu charakterisieren, als Überzeugung. Wenn Akzeptanz als nicht-propositional aufgefasst werden soll, fehlen Details dazu, wie genau nicht-propositionale Inhalte akzeptiert werden und welchen Akkuratheitsbedingungen Akzeptanz genügen müsste.

Natürlich widerlegt dies nicht das Argument hinter dem Problem der Propositionalität. Zwar haben Philosoph:innen, die diese Problematik gegen eine notwendige Überzeugungsbedingung für Verstehen anführen, meines Wissens keine Theorie nicht-propositionaler Akzeptanz präsentiert. Doch sicherlich ließe sich eine solche Theorie ausarbeiten; etwa anknüpfend an einen anderen Akzeptanzbegriff (z. B. van Fraassens Variante, die er im Kontext seines konstruktiven Empirismus entwickelt⁸²), an Überlegungen aus der Tugenderkenntnistheorie⁸³ oder zu visuellem Denken (etwa Kozak, 2023).

Andererseits: Wie problematisch ist das Problem der Propositionalität wirklich? Schließlich sind propositionale Einstellungen gegenüber nicht-propositionalen Dingen allgegenwärtig. Beispielsweise ist die Wirklichkeit vermutlich nicht propositional, und dennoch können wir propositionale Überzeugungen und Wissen darüber erwerben. Die Erklärung dieses Übergangs ist also keine Aufgabe, die sich lediglich im Kontext von Verstehen stellt.⁸⁴

Angesichts dessen ist es lohnenswert, bestehende Strategien zu betrachten, um den Übergang von nicht-propositionalen Repräsentationen zu propositionalen Einstellungen zu erklären. Wenn diese Strategien erfolgreich sind, besteht kein Bedarf für eine Alternative zu Überzeugung. Das wäre nicht zuletzt deshalb attraktiv, weil bereits diverse gut

⁸¹Eva Schmidt hat mich dankenswerterweise auf diesen Aspekt von Cohens Definition aufmerksam gemacht.

⁸²Van Fraassen ist Antirealist (siehe Fußnote 79) und nimmt an, dass wissenschaftliche Theorien nicht (näherungsweise) wahr, sondern lediglich empirisch adäquat sind. Daher erfordere die Akzeptanz einer Theorie bloß die Überzeugung, dass die Theorie empirisch adäquat ist (z. B. van Fraassen, 1980, S. 12). Allerdings spielt Verstehen in van Fraassens konstruktivem Empirismus keine Rolle (de Regt, 2015, S. 3786).

⁸³Die Tugenderkenntnistheorie geht davon aus, dass „nicht Überzeugungen, sondern vielmehr epistemische Akteure und ihre intellektuellen Tugenden primärer Gegenstand epistemischer Bewertung sind“ (Koppelberg, 2019, S. 79). Zagzebski schlägt daher vor, dass Tugenderkenntnistheoretiker:innen nicht darauf festgelegt seien, lediglich propositionale Objekte des Verstehens zuzulassen (Zagzebski, 2001, S. 248).

⁸⁴Dank gebührt abermals Eva Schmidt, die mich ermuntert hat, diesen wichtigen Punkt herauszuarbeiten.

ausgearbeitete philosophische Theorien zu Überzeugung existieren, an die Verstehenstheoretiker:innen anknüpfen können – zumal in den vorherigen Abschnitten deutlich wurde, dass Verstehen in vielen Fällen intuitiv mit Überzeugung einhergeht. Ich möchte daher zwei Strategien aus der philosophischen Literatur vorstellen, mithilfe derer sich erklären lässt, wie propositionale Einstellungen auf Grundlage visueller Repräsentationen erworben werden können.

Strategie 1: Manche visuellen Repräsentationen haben propositionale Inhalte

Die erste Strategie besteht darin, die Position zu relativieren, dass visuelle Repräsentationen keine propositionalen Inhalte haben. Betrachten wir zur Veranschaulichung ein konkretes Beispiel aus dem XAI-Kontext: Abbildung 7.1 zeigt eine visuelle Erklärung für die Klassifikation eines Bildes von Henry (einer Katze, die sich regelmäßig auf dem Campus der TU Dortmund herumtreibt) als ‚Bobtail‘, einer britischen Hunderasse. Der Standardauffassung zufolge ist ein Bild wie Abbildung 7.1 nicht wahr oder falsch, sondern mehr oder weniger akkurat – und zwar, weil visuelle Repräsentationen keine propositionalen Inhalte haben, die wahr oder falsch sein könnten.



Abbildung 7.1: Grad-CAM-Erklärung für die Klassifikation eines Fotos als ‚Bobtail‘ durch das Modell ResNet-50 (50,93%). Von rot hin zu blau eingefärbten Bildbereichen nimmt der Einfluss auf die Klassifikation ab. Bildbereiche mit keinem oder geringem Einfluss auf die Klassifikation sind aus Gründen der Übersichtlichkeit nicht eingefärbt. Bildquelle: Privat / eigene Darstellung unter Verwendung von Grad-CAM (Selvaraju et al., 2017). Siehe Teil IV für Details.

Ein prominenter Einwand gegen diese Standardauffassung stammt von Alex Grzankowski (2015). Sein Ausgangspunkt sind die expressiven Einschränkungen von Bildern. Im

Gegensatz zu den meisten Sprachen können Bilder logische oder explanatorische Relationen bestenfalls implizit ausdrücken (z. B. Negation, Konjunktion, Disjunktion, Modalität, Kontradiktion, Prädikation, Kausalität usw.; siehe dazu Crane, 2009; Gombrich, 1972; Sober, 1976). Grzankowski berücksichtigt dabei insbesondere die von Tim Crane (2009) und anderen diskutierte Unfähigkeit von Bildern, Aussagen zu negieren oder zu behaupten: Wenn ich schweigend das Foto von Kater Henry hochhalte, kann ich ohne zusätzliche Erläuterung nicht *versichern*, dass Henry im Foyer eines Universitätsgebäudes gesessen hat, oder *negieren*, dass dem so war. Ein anderes Beispiel: Ich kann mithilfe eines Bildes allein nicht ausdrücken, dass es Einhörner *nicht* gibt. Ich könnte natürlich ein Bild zeigen, auf dem keine Einhörner abgebildet sind (Grzankowski, 2015, S. 158). Damit würde ich aber wohl kaum aussagen, dass es keine Einhörner gibt – schließlich wäre das dann auch eine Aussage von Abbildung 7.1.

Grzankowski hält dagegen, dass man sich bejahend oder negierend auf Bilder beziehen kann. Wenn ich ein Foto von Henry hochhalte, wie er im Foyer des Universitätsgebäudes sitzt, kann ich zusätzlich so etwas sagen wie ‚So ist es gewesen!‘ oder ‚So war es nicht!‘. Offenkundig wird in einem solchen Fall irgendetwas negiert bzw. bejaht. Grzankowski argumentiert, dass dasjenige, das in einem solchen Fall negiert oder bejaht wird, der propositionale Inhalt des Bildes ist (ebd., S. 159–160).

Wenn das stimmte, ließe sich der Übergang von visuellen Repräsentationen zu propositionalen Einstellungen leicht erklären – die Inhalte der Überzeugungen, die in das Verstehen eines Bildes involviert sind, entsprechen den Inhalten dieses Bildes. Es besteht schlichtweg kein Bedarf für eine Alternative zu Überzeugung mehr. Allerdings hat diese Strategie längst nicht alle überzeugt (Korsmeyer, 1985, S. 264; Kozak, 2023, S. 37–39). Nur weil man sich durch propositionale Äußerungen auf etwas beziehen kann, folgt daraus noch nicht, dass das, worauf man sich bezieht, propositionalen Inhalt hat (ebd., S. 37). Wir beziehen uns andauernd in Form propositionaler Aussagen auf unsere materielle Umwelt, ohne dass diese eine propositionale Struktur hätte. Wenn man Grzankowskis Argument akzeptierte, stünde man damit eher alleine da.

Weit weniger kontrovers erscheint mir eine schwächere Annahme: Zumindest stark konventionalisierte visuelle Repräsentationen, wie Piktogramme oder Diagramme, haben propositionale Inhalte (Eaton, 1980; vgl. auch Korsmeyer, 1985, S. 258). Denn die grafischen Symbole, die solche Repräsentationen verwenden, um ihre Inhalte zu transportieren, lassen sich recht einfach durch sprachliche Symbole ersetzen – notfalls müssen dafür neue Ausdrücke eingeführt werden (Perini, 2005a, Fn. 5 auf S. 260). Beispielsweise argumentiert Laura Perini, dass Diagramme in Sprache übersetzt werden können: Bei einem Diagramm existieren Konventionen, die festlegen, welcher Sachverhalt bestehen muss, damit das Diagramm wahr ist (Perini, 2005b; vgl. auch Perini, 2005a, S. 259–260). Wenigstens für solche visuellen Repräsentationen lässt sich relativ unproblematisch annehmen, dass sie propositionale Inhalte haben. Der Grund ist, dass sie sich eines Symbolsystems bedienen, das ähnlich konventionalisiert ist wie sprachliche Repräsentationen. Die erste Strategie hilft aber nicht bei visuellen Repräsentationen, die weniger strikten Konventionen unterliegen,

wie zum Beispiel Fotos, Gemälden oder auch Salienzkarten. Hier greift die zweite Strategie.

Strategie 2: Nicht-propositionale Inhalte können in propositionale Strukturen übersetzt oder integriert werden

Die zweite Strategie besteht darin aufzuzeigen, dass die nicht-propositionalen Inhalte visueller Repräsentationen in Propositionen übersetzt oder integriert werden können. Offensichtlich können auch Bilder mit nicht-propositionalem Inhalt beschrieben und dadurch zumindest näherungsweise in propositionale (sprachliche) Repräsentationen übersetzt werden (siehe Crane, 2009, Kozak, 2023, S. 23–26). Romane schildern Orte und Menschen, Kunsthistoriker:innen beschreiben Gemälde und Skulpturen, ich habe oben bereits eine Beschreibung von Abbildung 7.1 skizziert usw. Diese Beispiele zeigen, dass die nicht-propositionalen Inhalte visueller Repräsentationen zu Inhalten propositionaler Einstellungen (wie Überzeugung) werden können, indem sie übersetzt werden. Die Inhalte der Salienzkarte aus Abbildung 7.1 ließen sich beispielsweise folgendermaßen übersetzen: ‚Der weiße, näherungsweise ovale Bereich in der Mitte des Bildes war ausschlaggebend für die Klassifikation‘ oder ‚Das System hat ein sitzendes Tier mit schwarz-weißem Fell erkannt‘.

Die entscheidende Frage ist natürlich, ob visuelle Repräsentationen *verlustfrei* in Sprache übersetzt werden können. Und das ist weniger eindeutig. Beispielsweise ist nicht offensichtlich, dass die Inhalte der Salienzkarte aus Abbildung 7.1 eine äquivalente sprachliche Entsprechung besitzen. Sobald man die visuelle Erklärung durch Aussagen wie ‚Der weiße, ovale Bereich in der Mitte des Bildes war ausschlaggebend für die Klassifikation‘ oder ‚Das System hat ein sitzendes Tier mit schwarz-weißem Fell erkannt‘ umschreibt, nimmt man eine Interpretation vor, die vom Inhalt der Salienzkarte abweicht. Das Bildklassifikationssystem verarbeitet keine Sprache, sondern Bilder (bzw. Pixelwerte). Ausschlaggebend war daher auch der hervorgehobene Bildbereich – kein ‚weißes Oval‘, und erst recht nicht Henrys ‚Fell‘ oder seine Körperhaltung. Zudem scheinen sich qualitative Eigenschaften wie Texturen oder Farbtöne nicht immer adäquat umschreiben zu lassen. Sollten die Inhalte visueller Repräsentationen wie jener aus Abbildung 7.1 kein linguistisches Äquivalent haben, ist eine verlustfreie Übersetzbarkeit in Propositionen fraglich.

Vielmehr geht die Mehrheit der Autor:innen davon aus, dass eine verlustfreie Übersetzung visueller Inhalte in Sprache nicht möglich ist (einen Überblick bietet Kjeldsen, 2015, S. 119). Und selbst wenn eine verlustfreie Übersetzung möglich wäre – die vollständige Umwandlung visueller in linguistische Repräsentationen ist ebenso schwierig wie umständlich (Korsmeyer, 1985) und scheint *de facto* nicht immer zu erfolgen. Wenn man eine visuelle Erklärung wie jene in Abbildung 7.1 betrachtet, formuliert man normalerweise keine inhaltlich äquivalente linguistische Erklärung. Man erfasst unmittelbar anhand des Bildes, warum das Foto als ‚Bobtail‘ klassifiziert worden ist.

Glücklicherweise gibt es einen Mittelweg zwischen Unübersetzbarkeit und perfekter Äquivalenz linguistischer und nicht-linguistischer Inhalte. Einige Autor:innen haben hervorgehoben, dass man sich durch deiktische Ausdrücke auf visuelle Repräsentationen

beziehen kann (siehe etwa Crane, 2009, S. 460–461; Kozak, 2023, S. 37–38). Dadurch können auch unübersetzbare nicht-linguistische Inhalte in propositionale Strukturen integriert werden: ‚Warum wurde das Foto als ‚Bobtail‘ klassifiziert?‘ – ‚Wegen *dieses* Bildbereichs.‘ Gleichmaßen kann man propositionale Einstellungen bezüglich Repräsentationen erwerben, die nicht-propositionale Inhalte haben, indem man sich deiktisch darauf bezieht. Ich kann zum Beispiel glauben, dass das Bild aufgrund von Merkmal ‚_____‘ als ‚Bobtail‘ klassifiziert worden ist, wobei die Lücke einem nicht-linguistischen Merkmal entspricht, auf das ich deiktisch verweisen kann. Diese Strategie bietet eine Antwort auf das Problem der Propositionalität, die nicht im Widerspruch zu der verbreiteten Annahme steht, dass zumindest manche visuellen Repräsentationen nicht-propositionale Inhalte haben.

7.3.2 Das Problem der Extraktion

Zusammengenommen formulieren die obigen Überlegungen gewissermaßen eine *Divide-et-impera*-Lösung für das Problem der Propositionalität. Nicht nur ist der Akzeptanzbegriff, wie er in der Verstehensdebatte verwendet wird, theoretisch nicht hinreichend ausgearbeitet worden, um das Problem zu lösen. Entscheidender ist, dass es verschiedene Möglichkeiten dafür gibt, den Übergang von nicht-propositionalen Verstehensobjekten zu propositionaler Überzeugung zu erklären. Erstens: Manche visuellen Repräsentationen haben propositionale Inhalte. Hier stellt sich das Problem der Propositionalität gar nicht erst. Zweitens: Die nicht-propositionalen Inhalte mancher visueller Repräsentationen können zumindest näherungsweise in Propositionen übersetzt werden. Diese Propositionen kommen dann als Inhalt propositionaler Einstellungen in Frage. Und drittens: Die nicht-propositionalen, unübersetzbaren Inhalte mancher visueller Repräsentationen können in propositionale Strukturen integriert werden, auch ohne sie zu übersetzen.

Meines Erachtens sind visuelle Repräsentationen in XAI-Kontexten nicht deswegen problematisch für Verstehen, weil ihre Inhalte nicht als Überzeugungsinhalte in Frage kommen. Ein zentrales Hindernis besteht vielmehr darin, die korrekte Extraktion propositionaler bzw. nicht-propositionaler Inhalte und deren Integration in Überzeugungen sicherzustellen. Denn viele visuelle (insbesondere piktoriale) Repräsentationen sind im Vergleich zu Sprache äußerst vage. John Kulvicki vergleicht Bilder mit kontextsensitiven Ausdrücken (z. B. ‚ich‘, ‚klein‘, ‚diese‘), die ihre Bedeutung je nach Kontext verändern (Kulvicki, 2020, S. 8–9): Dasselbe Bild kann in einem Kontext einen Kreis von schräg oben zeigen und in einem anderen eine Ellipse in Draufsicht (ebd., S. 9); Antoine de Saint-Exupérys berühmte Zeichnung eines Elefanten, der von einer Riesenschlange verschluckt wurde, zeigt in einem anderen Kontext lediglich einen Hut (vgl. Saint-Exupéry, 2023). Interpret:innen von Bildern brauchen daher Hintergrundwissen bzw. Kontextinformationen, um deren Aussage bestimmen zu können (Gombrich, 1972; Korsmeyer, 1985). Zwar ist auch Sprache mitunter kontextsensitiv und vage, doch weisen Bilder grundsätzlich eine größere „interpretative Offenheit“ auf (Kulvicki, 2020, S. 9; meine Übersetzung), welche die korrekte Extraktion

ihrer Inhalte erschwert. Nennen wir dies das *Problem der Extraktion*.⁸⁵

Das Problem der Extraktion ist ein zentraler Nachteil visueller Repräsentationen in XAI-Kontexten. Betrachten wir noch einmal Abbildung 7.1: Ohne weitere Erläuterung ist zum Beispiel unklar, dass es sich überhaupt um eine Erklärung handelt, was das Explanandum ist (die Klassifikation als ‚Bobtail‘, ‚Kran‘ oder die Position von Henrys bevorzugten Kraulstellen?), worin das Explanans besteht (waren abstrakte Kategorien wie *weiß*, *Fell*, *Pfote* entscheidend, oder Henrys spezifische Körperhaltung? Spielten die relativen räumlichen Relationen der verschiedenen Bildbereiche eine Rolle? Oder hatte alles davon einen Einfluss?) und welcher explanatorische Zusammenhang behauptet wird (kausal? gesetzmäßig? probabilistisch?). Sprache kann all diese Aspekte präzise ausdrücken, während Abbildungen aufgrund ihrer expressiven Einschränkungen nur nach korrekter Interpretation dazu in der Lage sind.⁸⁶ Angesichts dessen verwundert es nicht, dass viele Philosoph:innen Erklärungen ausschließlich als linguistisch bzw. propositional auffassen (siehe Fußnote 33).

Um dem Problem der Extraktion zu begegnen, sollten insbesondere visuelle Verstehensermöglicher durch geeignete Interpretationshilfen ergänzt werden, die dem Stakeholder dabei helfen, die relevanten Inhalte zu extrahieren und wahre Überzeugungen darüber auszubilden. In Kapitel 12 werde ich epistemisch vorteilhafte Verstehensermöglicher am Beispiel visueller XAI-Erklärungen näher charakterisieren. Vorerst genügt es festzuhalten, dass der nicht-propositionaler Charakter mancher Verstehensobjekte damit vereinbar ist, dass Verstehen Überzeugung (und nicht Akzeptanz, Verpflichtung o. ä.) erfordert.

Ich habe aber nicht bewiesen, dass Verstehen *ausschließlich* propositionale Einstellungen involviert. Gegner:innen meiner Position müssten zeigen, dass nicht nur Verstehensobjekte, sondern auch in Verstehen involvierte epistemische Einstellungen nicht-propositional sein können. Der Argumentation dieses Kapitels zufolge ist Überzeugung aber der derzeit attraktivste Kandidat für die in Verstehen involvierte epistemische Einstellung.

7.4 Instrumentelles Verstehen erfordert Überzeugung

Ich habe argumentiert, dass Verstehen *simpliciter* allem Anschein nach Überzeugung erfordert, und dass Akzeptanz keine einleuchtende Alternative dazu darstellt. Gilt dasselbe für instrumentelles Verstehen? Es gibt meines Erachtens keine stichhaltigen Gründe, die dagegen sprechen. Mir erscheint es äußerst unplausibel, dass jemand systembezogene Aspekte versteht, ohne entsprechende Überzeugungen zu besitzen.

Mithilfe der in Abschnitt 6.3 getroffenen Unterscheidung zwischen verschiedenen Repräsentationsebenen kann man zudem feinkörnig verschiedene Verstehensobjekte differenzieren und die zugehörigen Überzeugungen explizit machen. Betrachten wir das Beispiel

⁸⁵Das Problem der Extraktion ist nicht zu verwechseln mit Insa Lawlers (2021) ‚Extraktionsauffassung‘, die wir in Abschnitt 8.3.1 kennenlernen werden.

⁸⁶Vereinzelt wird diese Kritik auch von (informatischen) XAI-Forscher:innen geäußert, siehe etwa Achtibat et al., 2023; Rabold et al., 2020a,b; Rudin, 2019.

eines Empfehlungssystems, das eine Job-Bewerberin als geeignet eingestuft hat und dessen Einstufung mithilfe einer XAI-Methode erklärt wird. Ein Stakeholder könnte verstehen, warum das System die Bewerberin *laut der Erklärung* als geeignet eingestuft hat, ohne aber zu glauben, dass diese Erklärung wahr ist. Falls er die Erklärung hingegen für wahr hält, könnte der Stakeholder mit deren Hilfe aber auch verstehen, warum die Bewerberin *laut dem System* geeignet ist, ohne aber zu glauben, dass das System die Wirklichkeit akkurat genug abbildet. Wenn der Stakeholder durch die Verwendung des Empfehlungssystems einen Aspekt der Welt verstehen möchte, könnte er mithilfe des Systems (und einer Erklärung) verstehen, warum die Bewerberin *tatsächlich* geeignet ist. Doch egal, welche Repräsentationsebene man betrachtet: Instrumentelles Verstehen erfordert Überzeugung.

Wie ich in Abschnitt 7.3 vorgeschlagen habe, können auch auf Grundlage nicht-propositionaler Repräsentationen Überzeugungen ausgebildet werden. Bilder, Graphen, Diagramme oder Tabellen können einem Stakeholder ebenso wie linguistische Erklärungen epistemischen Zugang zu Informationen über relevante systembezogene Aspekte verschaffen. Beispielsweise zeigt eine Salienzkarte wie jene aus Abbildung 7.1 auf, dass bestimmte Pixel eines Bildes ausschlaggebend dafür waren, dass ein KI-System dieses Bild einer spezifischen Klasse zugewiesen hat. Auch solche potenziell nicht-propositionalen Informationen können Gegenstand von (propositionalen) Überzeugungen sein.

Wenn man den propositionalen Charakter der in Verstehen involvierten Einstellung wahrt, vereinfacht dies auch die Beantwortung von Fragestellungen der nächsten Subdebatte, die ich beleuchten möchte: jene zur Faktizität von Verstehen.

8 Faktizität von Verstehen

Ebenso, wie die meisten Philosoph:innen annehmen, dass Wissen Überzeugung erfordert, besteht auch große Einigkeit darüber, dass Wissen *wahre* Überzeugung voraussetzt, also faktiv ist: Wenn ich weiß, dass ein medizinisches Expertensystem eine Gastritis diagnostiziert hat, dann muss es der Fall sein, dass das fragliche System diese Diagnose getroffen hat. Das Verhältnis zwischen Verstehen und den Fakten ist hingegen komplizierter. Ein wichtiger Grund dafür ist, dass sich Verstehensobjekte von typischen Wissensobjekten unterscheiden. Während Wissen in paradigmatischen Fällen atomistisch ist und eine einzelne, simple Proposition betrifft, ist das Objekt von Verstehen umfangreicher und komplexer (siehe Abschnitt 6.1.3). Doch sobald wir es mit einem größeren „Informationsbrocken“ (Kvanvig, 2003, S. 192; meine Übersetzung) zu tun haben, steigt die Toleranz für falsche Überzeugungen (Malfatti, 2023, S. 47–48). Eine Reihe von Philosoph:innen ist daher zu dem Ergebnis gekommen, dass Verstehen nicht streng faktiv ist.

Wie die Lockerung des Faktizitäts-Kriteriums aussehen sollte, ist allerdings umstritten, und hängt nach Ansicht der meisten Autor:innen auch davon ab, ob es um propositionales, explanatorisches oder objektuales Verstehen geht (siehe Abschnitt 6.1). Ich stelle nachfolgend unterschiedliche Auffassungen hinsichtlich der Faktizität von Verstehen vor und argumentiere abschließend, dass instrumentelles Verstehen in einem kontextspezifischen Sinne moderat faktiv ist.

Um die Debatte zu strukturieren, ist es hilfreich, die verschiedenen Positionen auf einem Spektrum zwischen faktiv und nicht-faktiv anzuordnen. In Abschnitt 8.1 betrachte ich die streng nicht-faktive Position, derzufolge Verstehen nicht einmal voraussetzt, dass auch nur eine einzige Überzeugung wahr ist. In Abschnitt 8.2 betrachte ich die streng faktive Position am anderen Ende des Spektrums, die besagt, dass alle Überzeugungen, die Teil von Verstehen sind, wahr sein müssen. Zwischen diesen beiden Extremen liegen verschiedene Abstufungen moderat faktiver Positionen, die ich in Abschnitt 8.3 beleuchte. Schließlich stelle ich in Abschnitt 8.4 die moderat faktive Position vor, die ich in Bezug auf instrumentelles Verstehen vertrete.

8.1 Nicht-faktive Positionen

Dem nicht-faktiven Extrem des Spektrums zufolge erfordert Verstehen keinerlei wahre Überzeugungen. Nach meinem Wissen vertritt in dieser Form niemand diese Auffassung – mit guten Gründen. Ich gehe daher einige Positionen durch, die auf den ersten Blick

dem streng nicht-faktiven Ende des Spektrums anzugehören scheinen, und verdeutliche schließlich, weshalb eine solche Extremposition unplausibel ist.

Beginnen wir Linda Zagzebskis tugenderkenntnistheoretischer Position, derzufolge Verstehen eine Eigenschaft von Personen darstellt (Zagzebski, 2001, S. 245; siehe auch Fußnote 83). In diesem Zuge schlägt sie vor, dass Verstehen nicht durch traditionelle Begriffe aus der Wissensdebatte charakterisiert werden sollte, wie z. B. *Überzeugung*, *Wahrheit* oder *Proposition* (ebd., S. 245, 248). Damit fiele Verstehen aus dem hier zugrunde gelegten Spektrum heraus, da sich die Frage nach der Wahrheit von Überzeugungen gar nicht erst stellt.⁸⁷ Weiterhin deutet Zagzebski an, dass Verstehen rein internalistisch aufzufassen sei, während Wissen durch die Wahrheitsbedingung eine externalistische Komponente habe (ebd., S. 246).⁸⁸ Eine verstehende Person zeichne sich vielmehr durch Problemlösekompetenz aus, die lediglich eine indirekte Verknüpfung zu Wahrheit etabliere (ebd., S. 245). Ihr Vorschlag fügt sich daher besser in meine Diskussion von verstehensbezogenen Fähigkeiten ein (Kapitel 9).⁸⁹

Auch andere Autor:innen erwähnen eine internalistische Form des Verstehens. Stephen Grimm unterscheidet subjektives und objektives Verstehen, wobei subjektives Verstehen lediglich voraussetzt, dass das verstehende Subjekt eine Repräsentation begreift, die in ihr Weltbild passt, d. h. intern kohärent ist. Objektives Verstehen erfordert darüber hinaus auch, dass die Repräsentation (annähernd) wahr ist (Grimm, 2011, S. 91). In ähnlicher Form diskutiert Michael Strevens breites und enges Verstehen. Breites Verstehen involviert das Begreifen einer korrekten Erklärung. Enges Verstehen erfordert hingegen lediglich, dass die Erklärung in sich korrekt ist (Strevens, 2013, S. 512–513). Ich denke allerdings, dass sich diese Unterscheidungen besser durch die Differenzierung zwischen symbolischem und faktischem Verstehen einfangen lassen (siehe Abschnitt 6.2). Subjektives bzw. enges Verstehen ähneln demnach symbolischem Verstehen.

Eine weitere Position, die sich *prima facie* dem nicht-faktiven Lager zurechnen lässt, stammt von Catherine Elgin. Sie bezeichnet Verstehen sogar ausdrücklich als „non-factive“ (Elgin, 2007; Elgin, 2017, S. 33, 38). Innerhalb des hier zugrunde gelegten Spektrums der Faktizität ist diese Bezeichnung allerdings unpassend. Denn auch Elgin verlangt, dass Verstehen bis zu einem gewissen Grad mit den Fakten übereinstimmt: „The issue that divides factivists and non-factivists is not whether understanding must answer to the facts, but how it must do so“ (Elgin, 2007, S. 35). Zwar liegt Elgins Theorie recht weit auf der nicht-faktiven Seite, doch macht die genannte Übereinstimmung mit den Fakten ihre Theorie auf meinem Spektrum zu einer moderat faktiven Position. Ich betrachte ihre

⁸⁷ Gleiches gilt für Positionen, denen zufolge Verstehen keine propositionale Einstellung erfordert, siehe Kapitel 7.

⁸⁸ Ich komme in Kapitel 10 auf die Unterscheidung zwischen Internalismus und Externalismus zurück.

⁸⁹ Auch wenn verstehensbezogene Fähigkeiten eine wichtige Komponente meines Explikats darstellen, ähnelt meine Auffassung von Verstehen, so wie jene der meisten mit dem Thema befassten Erkenntnistheoretiker:innen, strukturell der klassischen Analyse von Wissen. Stefan Sleeuw kritisiert dieses Vorgehen und plädiert dafür, die Analyse von Verstehen grundsätzlich anders anzugehen (Sleeuw, 2023, S. 29).

Auffassung daher in Abschnitt 8.3.

Eine genuine nicht-faktive Position von Verstehen wäre in der Tat abwegig. Emma Gordon bietet dafür mehrere Argumente (Gordon, o. D., Kap. 2.b; Carter & Gordon, 2016, S. 425): Erstens korrigieren wir unsere Verstehenszuschreibungen gegenüber uns selbst oder einer anderen Person, wenn wir erfahren, dass relevante Überzeugungen falsch waren. Offenbar spielt die Wahrheit von Überzeugungen also eine Rolle für alltagssprachliche Zuschreibungen von Verstehen. Zweitens lässt eine nicht-faktive Position zu, dass zwei Personen, die jeweils konträre Überzeugungen bezüglich eines Themengebiets haben (Person A glaubt, dass p ; Person B glaubt, dass $\neg p$, usw.), gleichermaßen verstehen. Dieses Ergebnis ist äußerst kontraintuitiv. Drittens könnten auch in sich kohärente Hirngespinnste unter einer nicht-faktiven Auffassung von Verstehen als genuines Verstehen gelten. Unter Umständen müssten wir also sogar einer Anhängerin der *Flat Earth Society*, die glaubt, dass die Erde flach ist, geographisches Verstehen zuschreiben. Und viertens betrachten wir Verstehen üblicherweise als wertvolle kognitive Errungenschaft, als besonderen epistemischen Erfolg (z. B. Carter & Gordon, 2014; Kvanvig, 2003, 2009b; Pritchard, 2008, 2009). Ohne jegliche Verknüpfung zwischen den Überzeugungen eines verstehenden Subjekts und der Wahrheit lässt sich diese Einschätzung kaum erklären. Eine streng nicht-faktive Sichtweise, derzufolge Verstehen durch ausschließlich falsche Überzeugungen konstituiert sein könnte, scheidet also aus.

8.2 Faktive Positionen

Weit weniger abwegig wirkt die gegenteilige Extremposition, dass alle Überzeugungen, die Teil einer bestimmten Instanz von Verstehen sind, wahr sein müssen. Eine einzige falsche Überzeugung führt dieser Auffassung zufolge dazu, dass die fragliche Person nicht versteht. In Bezug auf atomistisch aufgefasstes propositionales Verstehen ist diese Position weitgehend konsensuell (Baumberger et al., 2017, S. 7; Elgin, 2007, S. 34): Plausiblerweise kann ich nicht verstehen, dass ein medizinisches Expertensystem eine Gastritis diagnostiziert hat, ohne dass es der Fall ist, dass das fragliche System diese Diagnose gestellt hat. Die Faktizität propositionalen Verstehens ist auch ein wichtiger Grund dafür, dass es in vielen Fällen mit Wissen gleichgesetzt wird (z. B. Baumberger, 2011, S. 87; Grimm, 2011, S. 85; Pritchard, 2009, S. 30). In Anbetracht meiner Überlegungen zu Verstehensobjekten (Abschnitt 6.1.3) denke ich allerdings, dass auch propositionales Verstehen nicht als streng faktiv aufgefasst werden sollte – ich komme gleich darauf zurück.

Auch in Bezug auf explanatorisches Verstehen, der am meisten diskutierten Unterart interrogativen Verstehens, vertreten viele Autor:innen einen strengen Faktivismus. Insbesondere innerhalb der Wissenschaftstheorie ist die Auffassung verbreitet, dass explanatorisches Verstehen dem Wissen einer Erklärung entspricht. Wie Peter Lipton es formuliert hat: „Understanding is not some sort of super-knowledge, but simply more knowledge: knowledge of the phenomenon and knowledge of its causal history“ (P. Lipton,

1990, S. 247). Kareem Khalifa charakterisiert diese Position daher als die vorherrschende Auffassung bezüglich Verstehen-warum (Khalifa, 2017, S. 16–18). Aber auch Autor:innen, die Verstehen nicht mit Wissen gleichsetzen, fassen explanatorisches Verstehen als faktiv auf. Beispielsweise schreibt Alison Hills: „You cannot understand why p if p isn't true, or if your belief as to why p is mistaken. So understanding why is factive [...]“ (Hills, 2016, S. 663). In ähnlicher Form konzipiert auch Duncan Pritchard explanatorisches Verstehen als faktiv: „[...] suppose that I believe that my house has burned down because of an act of vandalism, when it was in fact caused by faulty wiring. Do I understand why my house burned down? Clearly not“ (Pritchard, 2008, S. 333).

Sofern man annimmt, dass Erklärungen häufig komplexer sind, als das zumeist bemühte p weil q -Schema suggeriert, ist auch eine andere Lesart des strengen Faktivismus bezüglich explanatorischen Verstehens denkbar. Wie Christoph Baumberger et al. feststellen, könnte eine solche Position auch beinhalten, dass alle Propositionen (oder vielmehr: die Überzeugungen, die diese Propositionen zum Gegenstand haben) der Erklärung wahr sind (Baumberger et al., 2017, S. 7). Demzufolge würde eine Person, die eine einzelne falsche Überzeugung hinsichtlich einer Proposition hat, die Teil einer komplexen, umfangreichen (z. B. quantenmechanischen) Erklärung ist, keinerlei explanatorisches Verstehen besitzen. Sofern man – plausiblerweise – auch komplexe Erklärungen als Objekt explanatorischen Verstehens zulassen möchte, erscheint mir ein strenger Faktivismus weit weniger überzeugend, als es die simplen Beispiele von Hills oder Pritchard nahelegen.⁹⁰ Vielmehr spricht dies für einen moderaten Faktivismus in Bezug auf explanatorisches Verstehen.

Dasselbe gilt für propositionales Verstehen. In Abschnitt 6.1.3 habe ich vorgeschlagen, dass auch propositionales Verstehen nicht als atomistisch aufzufassen ist, sondern dass im Hintergrund ein komplexeres Verstehensobjekt steht. Wenn ich verstehe, dass ein medizinisches Expertensystem eine Gastritis diagnostiziert hat, dann speist sich diese Einsicht aus einem Überzeugungssystem, das diese simple Proposition betrifft (etwa hinsichtlich der Faktoren, die in die Diagnose eingeflossen sind). Andernfalls erscheint es überflüssig, von Verstehen (anstatt z. B. Wissen oder Überzeugung) zu sprechen. Dieses komplexere Überzeugungssystem kann plausiblerweise auch einzelne falsche Überzeugungen beinhalten, ohne dass ich aufhöre zu verstehen, dass das System eine bestimmte Diagnose getroffen hat.

Wenngleich nicht bezüglich propositionalen und explanatorischen Verstehens, so berufen sich wenigstens in Bezug auf objektuales Verstehen viele Autor:innen auf die Komplexität des Verstehensobjekts, um einen strengen Faktivismus zurückzuweisen. Zwar stimmen die meisten darin überein, dass falsche Überzeugungen den Verstehensgrad einer Person

⁹⁰Baumberger et al. (2017, S. 7) scheinen dem zuzustimmen; einen ähnlichen Punkt macht auch Andrés Páez (2019, S. 454). Julius Moravcsik weist zudem darauf hin, dass die Wahrheit von Komplexen mit p -weil- q -Struktur – anders als bei Strukturen wie $p \wedge q$ – graduell ist (Moravcsik, 1979, S. 203). Das spricht ebenfalls gegen einen strengen Faktivismus für explanatorisches Verstehen, da die Faktizität von Verstehensüberzeugungen – einschließlich ihrer Approximation der Wahrheit – eine plausible Dimension der Gradierbarkeit von Verstehen ausmacht (siehe Abschnitt 5.1).

reduzieren. Doch jemandem aufgrund einer einzigen falschen Überzeugung jegliches objektuale Verstehen, z. B. eines Themengebiets, abzusprechen, würde diese epistemische Errungenschaft zu einem selten erreichten Zustand machen. Wenn jemand beispielsweise glaubt, dass der Erste Weltkrieg an einem Mittwoch (anstatt Dienstag) begonnen hat, und abgesehen davon lediglich relevante wahre Überzeugungen über den Ersten Weltkrieg besitzt, würden wir dieser Person sicherlich einen hohen Grad an objektualem Verstehen zugestehen. Es ist wahrscheinlich, dass selbst eine Koryphäe auf dem Gebiet die ein oder andere falsche Überzeugung bezüglich des Ersten Weltkriegs aufweist. Solchen Personen jegliches Verstehen abzusprechen, nur weil sie einzelne falsche Propositionen glauben, erscheint äußerst unplausibel (Carter & Gordon, 2016, S. 425–426; Gordon, o. D., Kap. 2.c). Entsprechend sind streng faktive Auffassungen bezüglich objektualen Verstehens rar gesät (eine solche Position, jene von Insa Lawler, 2021, diskutiere ich in Abschnitt 8.3.1).

8.3 Moderat faktive Positionen

Wenn objektuales und, wie ich vorgeschlagen habe, auch propositionales sowie explanatorisches Verstehen nicht streng faktiv sind, und Nicht-Faktizität ebenfalls ausscheidet, stellt sich die Frage nach einem Mittelweg. Der prominenteste Vertreter eines solchen Mittelwegs ist vermutlich Jonathan Kvanvig mit seinem ‚Quasi-Faktivismus‘ (Kvanvig, 2003, 2009a; siehe auch Carter & Gordon, 2016; Mizrahi, 2012):

When the falsehoods are peripheral, we can ascribe understanding based on the rest of the information grasped that is true and contains no falsehoods. In such a case, the false beliefs are not part of the understanding the person has, even though they concern the very material regarding which the person has understanding. So in this way, the factive character of understanding can be preserved without having to say that a person with false beliefs about a subject matter can have no understanding of it. (Kvanvig, 2003, S. 201–202)

Kvanvigs Idee ist, dass objektuales Verstehen mit peripheren falschen Überzeugungen kompatibel ist, nicht aber mit falschen Überzeugungen, die zentrale Aspekte des jeweiligen Themengebiets betreffen (Kvanvig, 2009a, S. 341). Das eigentliche Verstehen ergibt sich aber stets aus den wahren Überzeugungen (insofern könnte man Kvanvig auch dem streng faktiven Lager zurechnen, mehr dazu in Abschnitt 8.3.2). Periphere falsche Überzeugungen sind lediglich kein Grund, einer Person jegliches Verstehen abzusprechen. Das Kriterium der Quasi-Faktizität verlangt also, dass alle zentralen Überzeugungen einer Instanz objektualen Verstehens wahr sind (ebd.). Wenn jemand fälschlicherweise glaubt, dass der Erste Weltkrieg an einem Mittwoch begonnen hat, mindert diese falsche Überzeugung das objektuale Verstehen dieser Person, beseitigt es aber nicht. Wenn jemand hingegen glaubt, dass der Erste Weltkrieg im Jahr 1618 begann, besitzt diese Person kein objektuales Verstehen des Ersten Weltkriegs, weil eine diesbezüglich zentrale Überzeugung falsch ist.

Ich finde es allerdings unplausibel, dass eine einzige zentrale falsche Überzeugung dazu führt, dass die betreffende Person *überhaupt kein* objektuales Verstehen besitzt (diese Intuition teilen u. a. auch Belkoniene, 2023, S. 60; Gordon, o. D., Kap. 2.c; und Riggs, 2009, S. 335–336). Zwar scheint eine große Anzahl falscher zentraler Überzeugungen irgendwann dazu zu führen, dass wir einer Person jegliches Verstehen absprechen. Mit einzelnen zentralen falschen Überzeugungen scheint Verstehen intuitiv allerdings vereinbar zu sein. Eine (in meiner Terminologie) moderat-faktive Position, derzufolge Verstehen auch mit zentralen falschen Überzeugungen kompatibel ist, vertritt Catherine Elgin. Hier sind zwei Beispiele, die sie diskutiert, um diese These zu stützen:

Evolution: „A second grader’s understanding of human evolution might include as a central strand the proposition that human beings descended from apes. A more sophisticated understanding has it that human beings and the other great apes descended from a common hominid ancestor (who was not, strictly speaking, an ape). The child’s opinion displays some grasp of evolution. It is clearly cognitively better than the belief that humans did not evolve. But it is not strictly true. Since it is central to her take on human evolution, [quasi-]factivists like Kvanvig must conclude that her take on human evolution does not qualify as understanding.“ (Elgin, 2007, S. 37)

Kopernikus: „A central tenet of Copernicus’s theory is the contention that the Earth travels around the sun in a circular orbit. Kepler improved on Copernicus by contending that the Earth’s orbit is not circular, but elliptical. [...] Despite the fact that Copernicus’s central claim was strictly false, the theory it belongs to constitutes a major advance in understanding over the Ptolemaic theory it replaced.“ (ebd., S. 37–38)

Ich denke, Elgin hat Recht damit, dass sowohl die Zweitklässlerin in *Evolution* als auch Kopernikus einen gewissen Grad objektualen Verstehens besitzen.⁹¹ Allerdings sollten wir an dieser Stelle zwischen falschen und annähernd wahren Überzeugungen unterscheiden. Eine der Subdimensionen der Faktizität von Verstehensüberzeugungen, die ich in Abschnitt 5.1.2 unterschieden habe, ist die Approximation der Wahrheit. Ich habe diese Subdimension durch den Vergleich zweier Personen illustriert: Person A glaubt, dass sich der *Homo erectus* vor genau 2,03 Millionen Jahren entwickelte und Person B glaubt, dass er sich vor exakt 2 Millionen Jahren entwickelte. Unter der Annahme, dass die Wahrheit ist,

⁹¹Kvanvig widerspricht dieser Einschätzung. Seiner Auffassung zufolge sind solche Verstehenszuschreibungen lediglich ‚honorierend‘ (engl. *honorific*), so wie wenn man vom aktuellen Stand des wissenschaftlichen Wissens spreche, obwohl man sich im Klaren sei, dass sich vieles davon künftig als falsch erweisen wird. Solche Fälle seien lediglich für pragmatische erkenntnistheoretische Untersuchungen von Interesse (Kvanvig, 2009a, S. 141–142). Ich denke hingegen, dass die Kompatibilität von Verstehen mit zentralen falschen oder nur annähernd wahren Überzeugungen erforderlich ist, um Intuitionen bezüglich verschiedener Verstehensgrade Rechnung zu tragen (siehe Abschnitt 5.1.2). So oder so: Instrumentelles Verstehen hat viele pragmatische Elemente, insofern sind Elgins Beispiele wenigstens für mein Vorhaben von Bedeutung.

dass der erste *Homo erectus* vor exakt 2,04 Millionen Jahren lebte, ist sowohl Bs als auch As Überzeugung streng genommen falsch. Dennoch würden wir beiden Personen Verstehen der menschlichen Evolution zugestehen, schließlich sind As und Bs Überzeugungen jeweils ziemlich nah an der Wahrheit. Analog schreiben wir den Protagonist:innen aus *Evolution* und *Kopernikus* deswegen Verstehen zu, weil die zentralen falschen Überzeugungen, die sie besitzen, nicht *völlig* falsch sind. Wie Elgin feststellt, glaubt die Zweitklässlerin aus *Evolution* schließlich nicht, dass Menschen gar nicht evolviert sind, oder dass sie sich aus Schmetterlingen entwickelt haben (ebd., S. 37).

Verstehen ist also intuitiv mit lediglich annähernd wahren zentralen Überzeugungen kompatibel. Doch was ist mit zentralen Überzeugungen, die völlig falsch sind, also die andere Subdimension der Faktizität betreffen? Adaptieren wir dafür *Evolution*, sodass die Zweitklässlerin eine eindeutig falsche, zentrale Überzeugung besitzt. Der Einfachheit halber erweitere ich das fragliche Objekt des Verstehens von der menschlichen Evolution auf jenes der Evolution im Allgemeinen:

Evolution*: Zweitklässlerin Jule besucht sowohl den Religions- als auch den Sachunterricht ihrer Grundschule. Sie setzt die Aussagen ihres Religionslehrers und ihrer Sachunterrichtslehrerin zu einem eher kruden Überzeugungssystem über die Evolution zusammen. Dazu gehören unter anderem folgende Überzeugungen: (1) Die heutigen Menschen sind nicht durch Evolution entstanden, sondern die Nachfahren von Eva und Adam. Gott hat Eva und Adam, die im Wesentlichen so waren wie die Menschen heute, im Laufe eines Tages erschaffen. (2) Die nicht-menschlichen Tiere, die man heute beobachten kann, sind aus einem sehr langwierigen Evolutionsprozess hervorgegangen. In dessen Verlauf haben sich die Vorfahren der heutigen nicht-menschlichen Tiere allmählich an ihren jeweiligen Lebensraum angepasst.

Jules Verstehen der Evolution enthält mit (2) eine (näherungsweise) wahre zentrale Überzeugung über die Evolution nicht-menschlicher Tiere (vielleicht hat sie nicht zugehört, als ihr Religionslehrer schilderte, wie Gott am vierten und fünften Tag die Tiere erschuf) und mit (1) eine eindeutig falsche zentrale Überzeugung über die Entstehung des Menschen. Ich denke, dass Jule die Evolution zu einem gewissen Grad versteht, obwohl eine der zentralen Überzeugungen falsch ist. Zwar gestehe ich gerne zu, dass ihr Verständnis dadurch signifikant reduziert wird. Ihr *keinerlei* Verstehen der Evolution zuzusprechen erscheint mir hingegen äußerst unplausibel. Wir können beispielsweise annehmen, dass Jule begriffen hat, wieso heutige Giraffen einen so langen Hals haben (weil unter ihren Vorfahren jene Tiere mit zufälligerweise längeren Hälsen relevante Nahrungsquellen erschließen konnten, dadurch einen Überlebensvorteil hatten und ihre langen Hälse daher an ihre Nachkommen vererben konnten). Sie hat also das Grundprinzip der Evolution verstanden und kann es auf nicht-menschliche Tiere anwenden – auch wenn sie Menschen vom Evolutionsprinzip ausschließt, was der Evolutionstheorie eindeutig widerspricht.

Kvanvigs Unterscheidung zwischen zentralen und peripheren Überzeugungen ist bei genauerer Betrachtung also nicht geeignet, um zu entscheiden, ob das Überzeugungssys-

tem einer Person faktiv genug ist, um als Verstehen gelten zu können. Doch kann diese Unterscheidung anderweitig fruchtbar gemacht werden. Erstens: In Abschnitt 5.1.2 habe ich vorgeschlagen, dass Verstehensgrade auch entlang der Faktizitätsdimension variieren können. Anknüpfend an Kvanvigs Unterscheidung könnte man die Position vertreten, dass je nachdem, welchen Stellenwert eine Überzeugung einnimmt, sie sich stärker oder schwächer auf den Verstehensgrad auswirkt. Eine zentrale falsche (oder nur näherungsweise wahre) Überzeugung mindert demnach den Verstehensgrad einer Person mehr, als wenn es sich um eine periphere Überzeugung gehandelt hätte. Verstehen wäre dann mit einer größeren Anzahl von peripheren denn zentralen falschen Überzeugungen kompatibel, da letztere den Verstehensgrad schneller so weit absenken können, dass wir nicht mehr von Verstehen sprechen würden. In Bezug auf Verstehen *simpliciter* halte ich diesen Vorschlag für plausibel. Zweitens: Bezüglich instrumentellen Verstehens bevorzuge ich eine andere Lesart der Zentral-peripher-Unterscheidung, die ich in Abschnitt 8.3.3 vorstelle. Dieser Vorschlag bewahrt Kvanvigs Idee, dass Verstehen die (annähernde) Wahrheit zentraler (nicht aber peripherer) Überzeugungen erfordert. Doch dazu später mehr.

Fassen wir noch einmal zusammen: Laut der zuvor skizzierten Argumentation ist Verstehen moderat faktiv, weil Verstehensobjekte stets eine gewisse Komplexität aufweisen. Eine streng faktive Position würde unrealistisch hohe Standards für Verstehen ansetzen und es zu einer außergewöhnlich seltenen epistemischen Errungenschaft machen. Normalerweise wird dieses Argument in Bezug auf objektuales Verstehen vorgebracht. Wie ich oben erläutert habe, denke ich allerdings, dass auch propositionales und explanatorisches Verstehen diese Schwelle der Komplexität überschreiten müssen, um noch als Verstehen zu gelten. Wenn das stimmt, ist Verstehen grundsätzlich als moderat faktiv aufzufassen. Allerdings ist noch nicht hinreichend klar geworden, worin dieser Faktizitätsgrad besteht.

In den folgenden Abschnitten möchte ich die moderate Faktizität von Verstehen daher weiter konkretisieren. Zunächst weise ich in Abschnitt 8.3.1 ein alternatives Argument für die Position zurück, dass Verstehen moderat faktiv ist. In Abschnitt 8.3.2 schlage ich vor, dass auch falsche Überzeugungen, die das Verstehensobjekt betreffen, einen Teil des Verstehens ausmachen. Schließlich plädiere ich in Abschnitt 8.3.3 dafür, dass man Kvanvigs Faktizitäts-Kriterium bewahren kann, wenn man die Zentralität von Verstehensüberzeugungen von kontextuellen Faktoren abhängig macht.

8.3.1 Moderate Faktizität und symbolisches Verstehen

Dem oben dargelegten Argument zufolge ist Verstehen moderat faktiv, weil die Komplexität von Verstehensobjekten eine streng faktive Position zu anspruchsvoll macht. Neben diesem – wie ich finde, überzeugenden – Argument findet sich in der Literatur noch ein zweites Argument für die moderate Faktizität von Verstehen, das häufig mit dem ersten vermengt wird. Da es immer wieder angeführt wird, möchte ich auch das zweite Argument kurz skizzieren und erläutern, warum ich mich nicht darauf stütze. Dessen Grundidee haben wir bereits in Abschnitt 7.2 kennen gelernt: Wissenschaftler:innen, d. h. paradigma-

tische Verstehenssubjekte, verwenden gängigerweise stark idealisierte oder anderweitig verfälschende Theorien und Modelle. Diese Theorien und Modelle weichen signifikant von der Wirklichkeit ab und sind, streng genommen, falsch. Nichtsdestotrotz erwerben Wissenschaftler:innen auf dieser Grundlage Verstehen von Aspekten der Welt. Daher, so die Schlussfolgerung, könne Verstehen nicht faktiv sein (Baumberger et al., 2017, S. 8; Elgin, 2007, 2017; Potochnik, 2015). Da bezüglich objektualem Verstehen weitgehender Konsens darüber herrscht, dass es nicht als streng faktiv aufgefasst werden sollte, findet sich dieses Argument vorwiegend in Bezug auf objektuales Verstehen. Doch Baumberger überträgt das Argument zusätzlich auch auf explanatorisches Verstehen: „We can only acknowledge that such [idealized] explanations provide some explanatory understanding if we admit that it is not always factive“ (Baumberger, 2011, S. 77).

Ich denke, dass dieses Argument die moderate Faktizität von Verstehen nicht zu etablieren vermag.⁹² In Abschnitt 7.2 habe ich anhand der Unterscheidung zwischen faktischem und symbolischem Verstehen aufgezeigt, dass die Verwendung inakkurater Modelle, Erklärungen oder Theorien nicht zwangsläufig dazu führt, dass die Nutzer:innen falsche Überzeugungen ausbilden. Zum einen können solche Repräsentationen rein symbolisch verstanden werden. Die resultierenden Überzeugungen betreffen dann in erster Linie die jeweilige Repräsentation und sind wahr, da sie implizit den Zusatz ‚Laut Repräsentation R...‘ beinhalten. Zum anderen kann man als kompetente Interpretin einer verfälschenden Repräsentation faktisches Verstehen erwerben, wenn man die Erklärung (o. ä.) nicht für uneingeschränkt wahr hält. Die Tatsache, dass Idealisierungen und andere Falschheiten zentral für die wissenschaftliche Praxis (und das Erklären komplexer KI-Systeme) sind, steht also nicht im Widerspruch dazu, dass Verstehen streng faktiv ist. Im Gegensatz zum ersten Argument kann das zweite Argument die moderate Faktizität von Verstehen nicht begründen, weil es auch mit einer streng faktiven Auffassung kompatibel ist.

Anstatt die Akkuratheit von für den Verstehenserwerb genutzten Repräsentationen und die Faktizität von Verstehen als zwei unabhängige Themen zu betrachten, werden beide innerhalb der Verstehensdebatte allerdings regelmäßig miteinander vermengt. Insa Lawler hat zwei hilfreiche Fragen formuliert, um beide Themen ausdrücklich voneinander abzugrenzen:

Relation question: How do the falsehoods (or the models that involve them) *relate to* the phenomenon in question? (Lawler, 2021, S. 6864; Hervorhebung im Original)

Content question: How do the falsehoods (or the models that involve them) *figure into the content* of understanding the phenomenon? (ebd., S. 6865; Hervorhebung im Original)

Die *Relationsfrage* betrifft die Faktizität der fraglichen Repräsentation, also ihr Verhältnis zur Wirklichkeit. Repräsentationen können ihr Repräsentandum mehr oder weniger

⁹²Ich habe diese Kritik selbst entwickelt, allerdings formulieren oder skizzieren auch andere Autor:innen ähnliche Einwände, vgl. Greco, 2014, S. 295–298; Hannon, 2021, S. 278–279; Hubert & Malfatti, 2023, S. 592–593; Kvanvig, 2009a, S. 342–343; Lawler, 2021; Malfatti, 2023, Kap. 2; van Camp, 2014, *passim*.

akkurat abbilden und entsprechend mehr oder weniger korrekt sein. Für die Beantwortung der Relationsfrage braucht man daher eine Theorie der Repräsentation oder Erklärung, mithilfe derer sich die Diskrepanz zur Wirklichkeit angeben lässt. Die *Inhaltsfrage* betrifft die Faktizität des Verstehens, also die Wahrheit der geglaubten Propositionen. Diese Frage lässt sich beantworten, indem man den Inhalt der Überzeugungen, die ein Subjekt auf Grundlage einer falschen Repräsentation ausbildet, betrachtet. Die Crux ist nun Folgendes: Wenn die Beantwortung der Relationsfrage aufdeckt, dass eine Repräsentation Falschheiten beinhaltet, folgt daraus noch nicht, dass die Antwort auf die Inhaltsfrage lauten muss, dass eine verstehende Person Falschheiten glaubt. Zwar determiniert die Antwort auf die Relationsfrage, was als Falschheit zu gelten hat (insofern sie von der Wirklichkeit abweicht) – und damit auch, welcher Inhalt einer Überzeugung als falsch zu gelten hätte, wenn er denn von einem Subjekt geglaubt werden würde. Doch wurde oben deutlich, dass inakkurate Repräsentationen nicht zwangsläufig falsche Überzeugungen zur Folge haben.

Lawler stellt der moderat faktiven Auffassung von Verstehen ihre ‚Extraktionsauffassung‘ (engl. *extraction view*) entgegen. Dieser Auffassung zufolge können bestimmte Falschheiten dabei helfen, relevante Informationen über reale Phänomene zu extrahieren (Lawler, 2021, S. 6876). Das mache manche Falschheiten epistemisch wertvoll (ebd., Kap. 7). Verstehen ist Lawler zufolge aber dennoch streng faktiv, weil diese Falschheiten nicht in falschen Überzeugungen resultieren, und damit die Inhaltsfrage nicht betreffen.

Ich denke, dass Lawler an dieser Stelle in die gleiche Falle tappt wie die Autor:innen, die sie kritisiert, und zwei verschiedene Fragen miteinander vermengt. Ich stimme ihr zu, dass die Extraktionsauffassung plausibel erklären kann, wie man auf Grundlage falscher Theorien, Erklärungen und Modelle zu wahren Überzeugungen hinsichtlich des Gegenstands dieser Repräsentationen kommen kann. Sie hat auch Recht damit, dass die Verwendung idealisierter oder anderweitig verfälschender Repräsentationen wenig damit zu tun hat, ob Verstehen faktiv ist. Was Lawler aber nicht zeigt, ist, dass Verstehen nicht mit falschen Überzeugungen kompatibel ist. Wie oben geschildert, argumentieren moderate Faktivist:innen, dass wir auch Personen mit peripheren und/oder zentralen falschen Überzeugungen bezüglich eines Themengebiets häufig Verstehen dieses Themengebiets zuschreiben. Die Frage ist also: Wenn beispielsweise ein Stakeholder eine XAI-Erklärung nicht gänzlich richtig interpretiert, und daher eine falsche Überzeugung ausbildet – besitzt dieser Stakeholder dann Verstehen? Moderate Faktivist:innen bejahen diese Frage. Lawler, die ausdrücklich eine faktive Auffassung von Verstehen vertritt (ebd., S. 6860), müsste diese Frage verneinen. Wie oben anhand des ersten Arguments dargelegt, ist eine solch streng faktive Auffassung aufgrund der Komplexität von Verstehensobjekten allerdings wenig plausibel.

8.3.2 Sind falsche Überzeugungen Teil des Verstehens?

Um meinen Einwand abzuwehren, könnte Lawler erwidern, dass der epistemische Zustand des Verstehens, den wir einer Person mit einzelnen falschen Überzeugungen zuschreiben, lediglich aus den *wahren* Überzeugungen besteht. Die falschen Überzeugungen sind dem-

nach nicht Teil des Verstehens. Wir erinnern uns, dass auch Jonathan Kvanvig eine solche Auffassung vertritt: „[...] the false beliefs are not part of the understanding the person has [...]. So in this way, the factive character of understanding can be preserved“ (Kvanvig, 2003, S. 201–202). Aus diesem Grund bezeichnet er seine Auffassung vermutlich auch als ‚quasi-faktiv‘ (Kvanvig, 2009a).

Es gibt mindestens zwei Argumente, die dagegen sprechen, dass die falschen Überzeugungen einer Person bezüglich des Verstehensobjekts einfach ausgeklammert werden können, wenn wir das Verstehen dieser Person bewerten. Das erste beruht auf der Prämisse, dass die falschen Überzeugungen unter Umständen mit den wahren Überzeugungen verwoben sind und letztere stützen. Dem zweiten Argument zufolge hängt die Beimessung eines Verstehensgrads auch von den falschen Überzeugungen einer Person ab. Ich betrachte nun nacheinander beide Argumente.

Das erste Argument, demzufolge die falschen Überzeugungen nicht einfach aus dem Verstehen einer Person herausgelöst werden können, formuliert Wayne Riggs anhand eines Beispiels:

Bootsunfall: „Suppose I understand my wife. I can reliably predict her behavior in a wide range of circumstances, I possess a lot of facts about her past and about her present dispositions, and so on. Included in this set of my beliefs about her past is that she suffered a very traumatic boating accident when she was 3 years old that resulted in her refusal ever to travel on water. [...] But, as it happens, she never had a boating accident. Instead, her parents neglectfully let her wander off while they were on vacation in Mexico when she was 2 (not 3), and she fell off a pier into the ocean. As she got older, her parents explained her fear of the water by making up a ‘boating accident’ that did not render them culpable for the resulting psychological damage. Obviously, my belief about her boating accident (as well as her belief) is false, and if understanding is strongly factive, it cannot be included in my understanding of my wife. [...] even if I was willing to grant that it should be ‘deleted’ from what constitutes my understanding of my wife, it is hard for me to imagine what it would be like to ‘subtract’ this one belief from my total understanding of her. What are we to do with all the evidential and explanatory connections that would be ‘left dangling’? Are they to be abandoned as well? If so, it looks as though my understanding of my wife suffers a terrible blow once we subtract my false belief about her alleged boating accident. But it does not seem plausible to me that such a judgment is reasonable.“ (Riggs, 2009, S. 335–336)

Verstehen, darin sind sich fast alle (auch Kvanvig) einig, hat mit dem Begreifen von Zusammenhängen zu tun (siehe Abschnitt 6.1.3 und Kapitel 9). Es handelt sich eben nicht nur um eine Reihe voneinander unabhängiger Überzeugungen, vielmehr erfasst eine verstehende Person, wie die einzelnen Elemente miteinander verknüpft sind – z. B. logisch, probabilistisch usw. Wenn man falsche Überzeugungen aus diesem Netzwerk entfernt, bleiben, wie Riggs richtigerweise feststellt, diverse ‚lose Enden‘ übrig (umso mehr, wenn

diese falschen Überzeugungen mit vielen anderen Überzeugungen verknüpft sind). Es erscheint schlichtweg unplausibel, dass Verstehen in diesem löchrigen Netzwerk besteht, wenn doch die falschen Überzeugungen dazu beitragen, es zu einem Netzwerk zu machen. Vielmehr sind auch die falschen Überzeugungen Teil dieses Netzwerks, und damit auch des Verstehens einer Person.

Für das zweite Argument können wir auf der von Riggs inspirierten Netzwerk-Metapher aufbauen. Die Idee ist, dass wir Verstehensgrade auch anhand der Faktizität bemessen (siehe Abschnitt 5.1.2). Denn plausiblerweise beeinflussen falsche Überzeugungen, die den Gegenstand des jeweiligen Verstehens betreffen, den Verstehensgrad einer Person. Ein Beispiel: Person A besitzt die wahren Überzeugungen *a*, *b* und *c* bezüglich Themengebiet *G*. Auch Person B glaubt wahrheitsgemäß, dass *a*, *b* und *c*. Zusätzlich glaubt B aber auch noch *x*, wobei *x* ebenfalls *G* betrifft, aber falsch ist. Intuitiv unterscheidet sich der Verstehensgrad von A und B. Vermutlich hat B einen geringeren Verstehensgrad als A, weil B mehr falsche Überzeugungen hinsichtlich *G* besitzt. Vielleicht ermöglicht die falsche Überzeugung *x* aber auch, dass B korrekte Verknüpfungen zwischen *a*, *b* und *c* herstellt, die Person A verschlossen bleiben, weil sie *x* nicht glaubt. Dadurch würde B im Vergleich zu A einen höheren Begreifensgrad besitzen. In diesem Fall könnte es sein, dass B insgesamt betrachtet sogar ein wenig mehr versteht als A.⁹³ So oder so ist unklar, wieso A und B unterschiedliche Verstehensgrade aufweisen, wenn die falschen Überzeugungen nicht Teil des Verstehens sind. (Diese Position ist damit vereinbar, dass wir Verstehen allein aufgrund der wahren Überzeugungen einen epistemischen Wert zuschreiben.)

Wir halten fest: Verstehen ist mit falschen Überzeugungen kompatibel und daher moderat faktiv, weil Verstehensobjekte stets komplex sind. Plausiblerweise macht es einen Unterschied, ob eine falsche Überzeugung zentral oder peripher ist. Wenn dem so ist, stellt sich die Frage, was eine Überzeugung zu einer zentralen bzw. peripheren Überzeugung macht.

8.3.3 Zentrale und periphere Überzeugungen

Kvanvig, von dem die Idee stammt, lässt weitestgehend offen, aufgrund welcher Eigenschaften eine Überzeugung als zentral oder peripher gelten muss. Einige andere Autor:innen haben aber versucht, seine Idee auszubuchstabieren. Dafür werden zwei verschiedene Strategien verfolgt: Die erste Strategie betrachtet die Eigenschaft einer Überzeugung, zentral oder peripher zu sein, als kontextunabhängig – nämlich als abhängig von der Stellung, welche die zugehörige Proposition innerhalb des verstandenen Themengebiets (o. ä.) einnimmt. Die zweite Strategie geht davon aus, dass der Kontext (mit-)bestimmt, welche Überzeugung in einem gegebenen Fall als zentral oder peripher zu gelten hat.

⁹³Catherine Elgin scheint solche Überzeugungen als ‚gelungene Falschheiten‘ (engl. *felicitous falsehoods*) anzusehen (Elgin, 2007, S. 41). Vornehmlich sind damit aber epistemisch vorteilhafte Abweichungen zwischen (wissenschaftlichen) Repräsentationen und ihren Repräsentanda gemeint (Elgin, 2007, 2017, 2022; siehe Abschnitt 7.2 und Kapitel 12).

Strategie 1: Die Zentralität von Verstehensüberzeugungen ist *unabhängig* vom Kontext

Der ersten Strategie zufolge ist eine Überzeugung, dass *p*, zentral oder peripher in Abhängigkeit davon, wie zentral oder peripher *p* innerhalb des ‚Informationsbrockens‘ ist, der verstanden wird. Manche Formulierungen von Kvanvig legen nahe, dass er diese Strategie verfolgt (etwa Kvanvig, 2009a, S. 341), und diese Auffassung klingt auch in *Evolution* und *Kopernikus* von Elgin sowie *Bootsunfall* von Riggs an. Zentrum und Peripherie sind demnach kontextunabhängige Eigenschaften der verstandenen Informationsmenge. Die Überzeugung, dass Menschen sich aus Affen entwickelt haben, ist zentral in Bezug auf die menschliche Evolution; die Überzeugung, dass die Planeten sich in perfekten Kreisbahnen um die Sonne bewegen, ist zentral für das kopernikanische Modell; und das vermeintliche Bootstrauma von Riggs’ Ehefrau ist zentral für ihre Persönlichkeit. An diese Lesart der Zentral-peripher-Unterscheidung habe ich auch im Erster-Weltkrieg-Beispiel appelliert: Der Wochentag, an dem der Erste Weltkrieg begann, ist peripher in Bezug auf das Themengebiet ‚Erster Weltkrieg‘, das Jahr (oder Jahrhundert) ist hingegen zentral.

Auch Adam Carter und Emma Gordon nehmen an, dass Verstehen moderat faktiv ist, und buchstabieren moderate Faktizität unter Berufung auf Kvanvig aus. In ihrem Aufsatz diskutieren die Autor:innen verschiedene Fallbeispiele, unter anderem adaptieren sie einen bekannten Fall von Jennifer Lackey (2008, S. 48). Lackeys ursprüngliches Beispiel soll zeigen, dass eine Hörerin Wissen durch das Zeugnis einer Person erwerben kann, die selbst (aufgrund fehlender Überzeugung) kein Wissen besitzt. Carter & Gordon argumentieren auf Grundlage ihrer Adaption von Lackeys Beispiel unter anderem, dass die Protagonistin kein Verstehen besitzt, weil ihr zentrale Überzeugungen fehlen. Daher lässt sich an dem Beispiel ablesen, welche Überzeugungen die Autor:innen als zentral erachten:

Kreationistische Professorin: „Stella is a devoutly Christian university professor, and her religious beliefs are grounded in a deep faith that she has had since she was a very young child. Part of this faith includes a belief in the truth of creationism and, accordingly, a belief in the falsity of evolutionary theory. Despite this, she fully recognizes that there is an overwhelming amount of scientific evidence against both of these beliefs. Stella, in fact, reads contemporary scientific journals and regularly teaches advanced graduate seminars on evolution, where students and colleagues alike admire the deep appreciation Stella has of how the theory holds together. Moreover, Stella readily admits that she is not basing her own commitment to creationism on evidence at all but, rather, on the personal faith that she has in an all-powerful Creator.“ (Carter & Gordon, 2016, S. 433)

Die Autor:innen vertreten die Auffassung, dass Stella aus *Kreationistische Professorin* die Evolutionstheorie nicht versteht, weil sie zentrale Annahmen der Evolutionstheorie für falsch hält – sie glaubt etwa, dass ‚Der moderne *Homo sapiens* entwickelte sich aus *Homo erectus*‘ falsch ist (ebd., S. 435). Wenn eine Theorie objektiv verstanden wird, erfordert dies

laut Carter & Gordon also nicht nur wahre Überzeugungen, welche die interne Logik der Theorie anbelangen, sondern auch wahre Überzeugungen hinsichtlich der Wahrheit bzw. Falschheit der Annahmen, die von der Theorie über ihren externen Gegenstand getroffen werden. Auch letztere Annahmen sind demnach zentral für das Verstehen einer Theorie.

Meines Erachtens ist *Kreationistische Professorin* hingegen ein klarer Fall von symbolischem Verstehen (siehe Abschnitt 6.2).⁹⁴ Stella versteht *symbolisch* die evolutionsbiologischen Zusammenhänge, die von der Theorie postuliert werden, bzw. die *laut der Theorie* wahr sind. Was Stella hingegen fehlt, ist *faktisches* Verstehen der Entstehung und Entwicklung biologischer Spezies. Sie versteht nicht, wie sich der *Homo sapiens* aus dem *Homo erectus* entwickelt hat, weil sie z. B. glaubt, dass alle modernen Menschen von Eva und Adam abstammen, die wiederum von Gott erschaffen wurden. Durch diese Trennung zwischen symbolischem und faktischem Verstehen lassen sich die zentralen falschen Überzeugungen, die Stella unterhält (z. B., dass der moderne *Homo sapiens* sich nicht aus dem *Homo erectus* entwickelt hat), in den Bereich des faktischen Verstehens auslagern. Die zentrale wahre Überzeugung, dass *laut der Evolutionstheorie* der moderne *Homo sapiens* aus dem *Homo erectus* entstanden ist, gehört hingegen zum symbolischen Verstehen. Und diese Überzeugung hat Stella.

Carter & Gordon würden einwenden, dass die zentralen Überzeugungen ohne den Zusatz ‚Laut der Evolutionstheorie...‘, die ich dem faktischen Verstehen zuordne, auch zentrale Überzeugungen symbolischen Verstehens sind (vgl. Carter & Gordon, 2016, S. 435–436). Wie die Autor:innen aber selbst einräumen, ist weitestgehend unklar, was eine Überzeugung zentral macht (ebd., S. 426; siehe auch Gordon, o. D., Kap. 2.c). Daher müssen wir uns mitunter auf Intuitionen verlassen. Und meines Erachtens ist es äußerst unplausibel, dass Stella, eine Person mit „deep appreciation [...] of how the theory holds together“ (siehe oben), kein objektuales Verstehen der Evolutionstheorie besitzt. Mein Vorschlag ist, dass symbolisches Verstehen nur symbolische Repräsentationen, wie z. B. Theorien, zum Gegenstand hat, und dass Überzeugungen, die das Repräsentandum dieser Repräsentationen betreffen, dafür nicht zentral sind.

Kreationistische Professorin macht deutlich, dass man zuerst das Objekt des Verstehens bestimmen muss, wenn man zentrale und periphere Überzeugungen unterscheiden will. Für symbolisches Verstehen sind andere Überzeugungen zentral als für faktisches Verstehen, weil das Verstehensobjekt ein anderes ist. Wir wissen aber immer noch nicht, wie sich zentrale und periphere Überzeugungen unterscheiden lassen, sobald das Objekt des Verstehens feststeht. Einen Vorschlag dazu findet man in Federica Malfattis Buch *Verstehen verstehen* (2023).

Malfatti geht davon aus, dass Verstehen unter anderem die Integration von Informationseinheiten in das noetische System einer Person erfordert. Das noetische System ist ein Netzwerk von Elementen, die vom verstehenden Subjekt geglaubt oder akzeptiert werden (ebd., S. 21–22). Ob ein Element zentral oder peripher ist, ergibt sich aus seiner Stellung

⁹⁴In einem späteren Aufsatz scheint auch Emma Gordon diese Lesart zu akzeptieren (Gordon, 2021, S. 4963).

innerhalb des noetischen Systems. Hängen viele andere Elemente von einem bestimmten Element ab, ist es zentral. Periphere Elemente sind hingegen weniger eng mit anderen Elementen verwoben. Würde eine Person aufhören, ein solches peripheres Element zu glauben oder zu akzeptieren, würde sich ihr noetisches System nur geringfügig verändern. Ein zentrales Element aufzugeben reißt hingegen, um im Bild zu bleiben, ein größeres Loch in das Netzwerk (ebd., S. 23, 49). Malfatti stellt also zunächst ein rein internalistisches Kriterium auf. Allerdings etabliert sie darüber hinaus eine externalistische Adäquatheitsbedingung, derzufolge Verstehen moderat faktiv ist (ebd., S. 43, 68). Demzufolge bildet das noetische System einer verstehenden Person die Wirklichkeit – sowohl in inhaltlicher als auch struktureller Hinsicht – größtenteils akkurat ab. Die Informationseinheiten eines noetischen Systems sind akkurat, wenn sie mit Tatsachen oder Ereignissen übereinstimmen; die Relationen zwischen diesen Einheiten sind akkurat, wenn sie den tatsächlichen Interdependenzen zwischen diesen Tatsachen bzw. Ereignissen entsprechen (ebd., S. 31). Die Zentralität innerhalb des noetischen Systems sollte idealerweise also mit der Zentralität innerhalb des verstandenen Phänomens oder Themengebiets einhergehen. Malfatti verfolgt demnach die erste Strategie und bestimmt die Zentralität von Verstehensüberzeugungen unabhängig von kontextuellen Faktoren.

Ein Vorschlag, der Malfattis Position ähnelt, stammt von Emma Gordon. Auch Gordon nimmt an, dass Verstehensüberzeugungen umso zentraler sind, je mehr andere Überzeugungen davon abhängen (Gordon, 2021, S. 4965). Denn aus diesen Abhängigkeiten ergebe sich, welche Überzeugungen über das verstandene Thema *praktisch relevant* seien:

Practical Relevance: S's belief B is (highly) practically relevant to S's objectual understanding, $\varphi = \text{Df.}$ S believes B, and were S to believe falsely whether B, this would significantly undermine S's reliability when S makes predictions or manipulates information with respect to subject matter φ .

Centrality: S's belief B is central to S's objectual understanding, $\varphi = \text{Df.}$ B is highly practically relevant to S's understanding of φ . (ebd.; Hervorhebung im Original)

Gordon definiert Zentralität also durch praktische Relevanz, und praktische Relevanz ergibt sich aus der Fähigkeit der Vorhersage und der Manipulation von themenbezogenen Informationen (ich komme in Abschnitt 9.2 auf diese Fähigkeiten zurück). Solange eine Überzeugung keine signifikanten negativen Auswirkungen auf diese Fähigkeiten hat, ist sie demnach peripher. Gordons Vorschlag ist also praktisch orientiert, aber nicht kontextsensitiv.

Im Gegensatz dazu argumentiert eine Reihe von Autor:innen, dass Verstehenszuschreibungen davon abhängen, ob ein Verstehenssubjekt diejenigen Dinge tun bzw. Probleme lösen kann, die *im jeweiligen Kontext* relevant sind. Diese Überlegungen dienen zwar nicht der Identifizierung zentraler Verstehensüberzeugungen, sprechen aber gegen eine kontextunabhängige Auffassung verstehensbezogener Fähigkeiten (in Abschnitt 9.3 vertrete auch ich die Position, dass der Kontext beeinflusst, welche Fähigkeiten instrumentelles Verstehen erfordert). Beispielsweise definiert Christoph Kelp absolutes Verstehen durch den Grad der

Annäherung an maximales Verstehen, der die erfolgreiche Bewältigung kontextabhängiger Aufgaben ermöglicht (Kelp, 2015, S. 3813). Henk de Regts Intelligibilität (engl. *intelligibility*) geht ebenfalls in diese Richtung. In *Understanding Scientific Understanding* definiert er Intelligibilität als Cluster von Eigenschaften einer Theorie, das es spezifischen Wissenschaftler:innen in einem kontextspezifischen Sinne ermöglicht, diese Theorie zu nutzen (de Regt, 2017, S. 40). Außerdem fügt sich hier der in Abschnitt 6.1.2 skizzierte Vorschlag von Michael Stuart ein, praktisches Verstehen unter anderem durch das Verfügen über zielbezogene Fertigkeiten zu definieren (Stuart, 2025). Diese Überlegungen bringen uns zur zweiten Strategie, derzufolge kontextuelle Faktoren berücksichtigt werden müssen, um zentrale und periphere Überzeugungen voneinander zu unterscheiden.

Strategie 2: Die Zentralität von Verstehensüberzeugungen ist *abhängig* vom Kontext

Der zweiten Strategie zufolge beeinflussen kontextuelle Faktoren, was eine Überzeugung zentral oder peripher macht. Daniel Wilkenfeld (2015, Manuskript) verfolgt diese Strategie.⁹⁵ Er schlägt vor, dass die Kontextsensitivität insbesondere dann augenscheinlich wird, wenn Verstehen von praktischer Bedeutung ist. Hier ist sein Beispiel:

Metall: „[...] consider two scientists working on the same metal, who form the same beliefs about it—a set of true beliefs about its conductive properties and a false belief about its physical strength. However, scientist A is trying to use it to build a bridge, and scientist B is trying to use it to make a wire. We contend that, though they believe all the same things, we would and should attribute understanding to the latter but not the former.“ (ebd., o. S.)

In diesem Beispiel hängt die Zentralität einer Überzeugung, die Teil des objektualen Verstehens des fraglichen Metalls ist, davon ab, was das verstehende Subjekt mit seinem Verständnis anfangen will. Da die falsche Überzeugung bezüglich der Stabilität des Metalls lediglich As Vorhaben, eine Brücke zu bauen, beeinträchtigt, versteht Wissenschaftler:in B das Metall, nicht aber A. Für A ist die falsche Überzeugung also zentral, für B nicht – darum sprechen wir nur B Verstehen zu. Wilkenfelds Idee lautet also, dass der Kontext maßgeblich beeinflusst, welche Überzeugung als zentral zu gelten hat. Trotzdem lassen sich Verstehensgrade kontextübergreifend unterscheiden: Wenn die Kontexte, in denen Person A versteht, eine echte Teilmenge der Kontexte ist, in denen B versteht, dann ist Bs Verstehen besser als jenes von A (ebd., o. S.). Wilkenfeld bezeichnet seine Position als ‚kontextuelle Quasi-Faktizität objektualen Verstehens‘:

⁹⁵Vielen Dank an Daniel Wilkenfeld, dass er mir das (unveröffentlichte) Manuskript zur Verfügung gestellt und zugestimmt hat, dass ich mich darauf beziehe. Ähnliche Überlegungen wie jene von Wilkenfeld finden sich auch bei John Bengson, 2018, S. 36. Einen Vorschlag für die kontextsensitive Bestimmung der Zentralität von Überzeugungen explanatorischen Verstehens macht Miloud Belkoniene, 2023, Kap. 4.

The Contextual Quasi-Factivity (CQF) of objectual understanding: one understands object *o* only if one's central beliefs about *o* are true, where a belief's centrality is pragmatically determined. (ebd., o. S.)

Verdeutlichen wir Wilkenfelds Idee anhand der zuvor betrachteten Fallbeispiele – einige davon diskutiert Wilkenfeld auch selbst. Ob die Zweitklässlerin aus *Evolution* die Evolution versteht, hängt auch davon ab, was sie mit ihrem Verstehen anfangen will: Wenn sie erklären will, wie sich kleine evolutionäre Anpassungen zu großen Veränderungen aufsummieren können, besitzt sie Verstehen. Wenn sie die menschliche Entwicklung im Besonderen darstellen möchte, versteht sie hingegen nicht (ebd., o. S.). Wenn Jule aus *Evolution** die menschliche Evolution erklären möchte, ist ihre falsche Überzeugung zentral. Wenn es hingegen um die Evolution der Giraffe oder eines anderen nicht-menschlichen Tieres geht, ist sie peripher. Gleichmaßen taugt Kopernikus' Annahme der perfekten Kreisbahnen dafür, den grundlegenden Aufbau unseres Sonnensystems zu verdeutlichen. Jemand, der die genaue Position des Mars vorhersagen möchte, besitzt hingegen kein Verstehen, wenn er glaubt, dass die Planeten sich auf perfekten Kreisbahnen bewegen. Wilkenfeld bezieht sich auch auf *Bootsunfall* von Riggs: Im Alltag spiele es keine große Rolle, wie genau die Angst von Riggs' Ehefrau vor der Fortbewegung zu Wasser entstanden ist. Riggs versteht seine Frau also. Wollte er ihre Phobie aber durch Psychoanalyse kurieren, wäre die falsche Überzeugung kontextuell zentral – er würde nicht verstehen (ebd., o. S.).

Wilkenfelds Strategie kann auch Intuitionen erklären, denen zufolge andere Kontextfaktoren Verstehenszuschreibungen beeinflussen. Beispielsweise sind wir gewillt, einer Zweitklässlerin Verstehen der Evolution zuzuschreiben, wenn sie glaubt, dass sich Menschen aus Affen entwickelt haben. Einer Professorin der Evolutionsbiologie würden wir aber womöglich jegliches Verstehen der Evolution absprechen, wenn sie diese Überzeugung unterhalten würde (vgl. auch Bachmann, 2020, S. 79–80; Kelp, 2015, S. 3803; Kelp, 2017, S. 254; Khalifa, 2017, S. 5; Søndergaard, 2023, S. 2609; Wilkenfeld, 2013, Kap. 3.2). Der kontextuellen Quasi-Faktizität zufolge sind für eine Zweitklässlerin andere Überzeugungen zentral als für eine Universitätsprofessorin, entsprechend versteht die erstere, nicht aber die letztere.

Im folgenden Abschnitt schlage ich vor, dass eine kontextunabhängige Bestimmung der Zentralität von Überzeugungen für Verstehen *simpliciter* zwar plausibel ist, dass die Zentralität einer *instrumentellen* Verstehensüberzeugung aber vom Kontext abhängt.

8.4 Instrumentelles Verstehen ist moderat faktiv

Instrumentelles Verstehen ist moderat faktiv: Eine streng faktive Position formuliert bereits in alltäglichen Kontexten kaum erreichbare Standards, da Verstehensobjekte stets eine gewisse Komplexität aufweisen. Das gilt umso mehr für XAI-Kontexte, in denen auch Personen mit wenig Hintergrundwissen äußerst komplexe KI-Systeme (oder einzelne ihrer Aspekte) verstehen sollen. Eine streng nicht-faktive Auffassung instrumentellen

Verstehens scheidet ebenfalls aus. Wenn jemand keine einzige (wenigstens annähernd) wahre Überzeugung bezüglich des Verstehensobjekts besitzt, ist unklar, weshalb es sich dabei überhaupt um Verstehen handeln sollte. Bleibt noch eine moderat faktive Position, die mit intuitiven Verstehenszuschreibungen übereinstimmt, ohne kaum erreichbare Standards zu setzen.

Moderate Faktizität lässt sich durch eine Unterscheidung zwischen zentralen und peripheren Überzeugungen konkretisieren, die wiederum auf kontextsensitive und kontextunabhängige Weise bestimmt werden können. In Bezug auf Verstehen *simpliciter* haben beide Positionen etwas für sich. Verstehenszuschreibungen sind in vielen Fällen kontextsensitiv, und es ist attraktiv, diese Kontextsensitivität mit der Zentralität von Überzeugungen zu verknüpfen. Zudem könnten Befürworter:innen eines kontextsensitiven Ansatzes meinen mithilfe von *Evolution** formulierten Einwand (Abschnitt 8.3) zurückweisen. Wir erinnern uns: Kvanvig zufolge wird Verstehen durch zentrale falsche Überzeugungen verhindert. Anhand von *Evolution** habe ich gezeigt, dass dies nicht immer der Fall ist. Mithilfe eines kontextsensitiven Ansatzes ließe sich dieser Einwand abwehren: Man könnte behaupten, dass wir bei *Evolution** intuitiv einen Kontext annehmen, in dem die falsche Überzeugung über die menschliche Evolution kontextuell peripher ist – die Fallbeschreibung böte keinen Anlass anzunehmen, dass die Erreichung von Jules Zielen gefährdet wird, wenn sie glaubt, dass der moderne Mensch von Eva und Adam abstammt. Wenn man die Zentralität von Verstehensüberzeugungen vom Kontext abhängig macht, ließe sich auch angesichts von Fällen wie *Evolution** Kvanvigs Vorschlag aufrechterhalten, dass zentrale falsche Überzeugungen Verstehen verhindern (und nicht bloß den Verstehensgrad reduzieren).

Doch ist es wirklich überzeugend anzunehmen, dass wir bei jeder noch so alltäglichen Verstehenszuschreibung die Ziele der betreffenden Person im Blick haben (und notfalls Ziele unterstellen)? Zwar mag es Kontexte geben, in denen (z. B.) eine wahre Überzeugung über den Wochentag, an dem der Erste Weltkrieg begann, wichtiger ist als eine über das Jahrhundert. Gleichzeitig scheint es auch eine relevante kontextunabhängige Redeweise von Verstehen zu geben, derzufolge der fragliche Wochentag nur ein peripheres Detail des Kriegsgeschehens ausmacht. Vorschläge wie jene von Gordon oder Malfatti, die tatsächlichen Verknüpfungen ‚in der Welt‘ mit jenen zwischen Überzeugungen abzugleichen, können diese Intuition gut einfangen. Zwar müsste man dann ggf. von Kvanvigs Annahme abrücken, dass falsche zentrale Überzeugungen Verstehen zwangsläufig verhindern. Aber man könnte daran festhalten, dass in einem (wie auch immer bestimmten) kontextunabhängigen Sinne zentrale Überzeugungen den Verstehensgrad stärker reduzieren als falsche Überzeugungen über nebensächliche Aspekte.

Ich denke daher, dass viele Fälle von Verstehen *simpliciter* durch eine kontextunabhängige Bestimmung der Zentralität von Verstehensüberzeugungen adäquat beschrieben werden können. Aufgrund seines zweckgerichteten Charakters gilt dies aber nicht für instrumentelles Verstehen. Anders als bei Verstehen *simpliciter* steht bei instrumentellem Verstehen immer im Vordergrund, was der fragliche Stakeholder mit seinem Verständnis anfangen will (nämlich die Erfüllung des jeweils relevanten Desideratums). Es handelt sich um eine

hochgradig pragmatische Verstehensform. Für instrumentelles Verstehen folge ich daher der zweiten Strategie, die Zentralität von Überzeugungen durch den Kontext zu bestimmen. Instrumentelles Verstehen ist demzufolge mit falschen Überzeugungen kompatibel, solange diese kontextuell peripher sind. Kontextuell zentrale Überzeugungen müssen hingegen wahr oder wenigstens näherungsweise wahr sein.

Doch welche Kontextfaktoren wirken sich auf die Zentralität einer instrumentellen Verstehensüberzeugung aus? Ich möchte nun vorschlagen, dass das Desideratum des fraglichen Stakeholders den wichtigsten Kontextfaktor darstellt. Das dürfte kaum überraschen: Die Erfüllung gesellschaftlicher Desiderata ist das Ziel instrumentellen Verstehens. Jene Überzeugungen, die dazu beitragen, dieses Ziel zu erreichen, sind daher zentral und müssen zumindest annähernd wahr sein. Im Gegensatz zur kontextunabhängigen Perspektive lässt sich die Zentralität von instrumentellen Verstehensüberzeugungen ohne Berücksichtigung des Desideratums nicht bestimmen. Dies lässt sich anhand eines Beispiels verdeutlichen:

Bewerbung:⁹⁶ Programmiererin Lina ist die Cheftwicklerin eines Empfehlungssystems namens *JobAI*, das bei der Personalauswahl eingesetzt wird. In das KI-System werden Informationen aus den Bewerbungsunterlagen der jeweiligen Bewerber:innen eingespeist – zum Beispiel der Lebenslauf, Zertifikate über zusätzliche Qualifikationsmaßnahmen oder Arbeitszeugnisse. Anhand dieser Informationen ordnet *JobAI* allen Bewerber:innen jeweils einen Wert entsprechend der vom System abgeschätzten Eignung zu. Der Personalmanager Fritz, der in einem anderen Unternehmen arbeitet, verwendet dieses System, um geeignete Kandidat:innen für eine zu besetzende Stelle auszuwählen. *JobAI* vergibt an Bewerberin Magda nur 3 von 10 möglichen Punkten, während Fritz Magda als besonders vielversprechende Kandidatin einschätzt.

Nehmen wir an, dass für Lina und Fritz jeweils eine andere Variante des Fairness-Desideratums im Vordergrund steht. Fritz muss herausfinden, ob die Bewertung von Magda durch *JobAI* fair ist. Bei KI-Systemen, die mit historischen Daten zu erfolgreichen bzw. erfolglosen Bewerbungen trainiert werden, besteht die Gefahr, dass sie männliche Bewerber systematisch bevorzugen: Da in vielen Berufen Männer statistisch gesehen häufiger eingestellt wurden, kann das Geschlecht (oder ein Proxy dafür) bei historischen Daten bis zu einem gewissen Grad vorhersagen, ob jemand für eine vakante Stelle ausgewählt wurde. KI-Systeme können dieses diskriminierende Muster dann heranziehen, um die vorgebliche Eignung von Bewerber:innen abzuschätzen.⁹⁷

Fritz ist bewusst, dass manche KI-Systeme solche Verzerrungen aufweisen und dadurch z. B. Frauen statistisch benachteiligen. Angesichts dessen muss er die Gründe oder Ursachen für die Bewertung von Bewerberin Magda kennen, um überprüfen zu können, ob diese

⁹⁶Das Beispiel ist angelehnt an ähnliche Beispielfälle von Kevin Baum et al. (2022) und Timo Speith et al. (2024).

⁹⁷Im Jahr 2018 machte z. B. die Meldung Schlagzeilen, dass der von Amazon verwendete Algorithmus zur Bewertung von Jobbewerbungen männliche Bewerber bevorzugte (Dastin, 2018).

zu einer unfairen Bewertung geführt haben. Mithilfe der Erklärung einer zuverlässigen XAI-Methode könnte er wahre Überzeugungen darüber ausbilden, welche Merkmale von Magdas Bewerbungsunterlagen (etwa die universitäre Abschlussnote oder das Französischniveau) zu der schlechten Bewertung geführt haben. Diese Überzeugungen sind zentral, um instrumentell zu verstehen, ob die Entscheidung von *JobAI* fair ist.

Betrachten wir als nächstes Lina. Sie ist Chefentwicklerin von *JobAI* und wertet die Daten aus Fritz' Unternehmen aus, um das System zu verbessern. Überzeugungen bezüglich der in Magdas Fall ausschlaggebenden Merkmale sind für Lina peripher oder sogar irrelevant. Um zu verstehen, ob das System fair ist, braucht sie wahre Überzeugungen darüber, ob Frauen statistisch gesehen als weniger geeignet eingestuft werden, oder welche Merkmale der Eingabedaten häufig für die Einstufung weiblicher Bewerber:innen herangezogen werden. Mehr noch, als Systementwicklerin will Lina nicht bloß herausfinden, ob das System fair ist. Sie will auch verstehen, wie sie das System fairer machen kann, falls das Desideratum substantiell unerfüllt ist. Wenn sich herausstellt, dass *JobAI* tatsächlich Frauen benachteiligt, braucht sie entsprechend auch wahre Überzeugungen bezüglich der Ursachen dieser Ungleichbehandlung. Besteht das Problem aufgrund eines historischen Bias in den Trainingsdaten, waren die Daten nicht repräsentativ für die relevante Gruppe oder wurden sie vielleicht unsachgemäß erhoben bzw. für das Training aufbereitet? Wahre Überzeugungen bezüglich der Ursache des Problems sind für Linas instrumentelles Verstehen zentral. Für Fritz spielen sie hingegen keine Rolle, um sein Fairness-Desideratum (epistemisch) zu erfüllen.

Das Beispiel von Lina und Fritz aus *Bewerbung* zeigt, dass wie auch in *Metall* das konkrete Ziel beeinflusst, welche Überzeugungen wahr oder wenigstens annähernd wahr sein müssen, damit eine Person instrumentell versteht. In den für mich relevanten XAI-Kontexten ist dieses Ziel die epistemische, mitunter auch substantielle Erfüllung eines bestimmten Desideratums. Überzeugungen über systembezogene Aspekte, die sich darauf auswirken, sind daher zentral.

Allerdings wäre es zu viel verlangt, wahre Überzeugungen über *alle* systembezogenen Aspekte zu erwerben, welche die Erfüllung eines Desideratums betreffen. Betrachten wir dazu noch einmal das *Bewerbung*-Beispiel und nehmen an, dass *JobAI* Bewerber:innen aller Geschlechter fair beurteilen soll, wenn es geeignete Kandidat:innen für eine vakante Stelle auswählt. Dieses Desideratum berührt eine Vielzahl komplexer Themengebiete: Die Einschätzung der Fairness von Bewerbungsprozessen schließt unter anderem ethische, politische, psychologische, historische, feministische und sozialwissenschaftliche Erwägungen ein. Durch den Einbezug von KI kommen diverse informatische und mathematische Aspekte hinzu. Zudem beeinflusst auch die Anwendungsdomäne, wann die Auswahl von Bewerber:innen als fair zu gelten hat – in einer Branche, in der Frauen aus kontingenten Gründen überrepräsentiert sind (etwa der Kinderbetreuung) gelten womöglich andere Kriterien als z. B. in einem Frauenhaus. Viele weitere systembezogene Aspekte ließen sich nennen. Kein einzelnes Individuum kann sich realistischweise ein hochgradiges Verständnis all dieser Aspekte erarbeiten. Das würde mehr Zeit, kognitive Kapazitäten und

vielleicht auch eine größere Vielfalt an Talenten erfordern, als der Einzelnen zur Verfügung stehen (Hardwig, 1985).

Die Frage ist also, wie sich die Menge zentraler Überzeugungen auf ein realistisches Maß einschränken lässt. Ich vermute, dass es darauf keine allgemeingültige Antwort gibt – XAI-Kontexte und die Erfordernisse der dort relevanten Desiderata sind dafür schlichtweg zu vielfältig. Anstatt einer klaren Regel möchte ich daher eine Heuristik vorschlagen, mithilfe derer die für instrumentelles Verstehen zentralen Überzeugungen angemessen limitiert werden können. Dafür müssen wir uns vergegenwärtigen, dass die Erlangung epistemischer Güter auch in XAI-Kontexten häufig auf epistemischer Arbeitsteilung beruht (siehe Abschnitt 2.3.3). Beispielsweise stützen sich die Entwickler:innen eines KI-Systems auf die Erkenntnisse von Generationen an Informatiker:innen vor ihnen, auf Personen, die Trainingsdaten zusammengestellt und annotiert haben oder auf Aussagen von Expert:innen bezüglich der Anwendungsdomäne, in der das System eingesetzt werden soll. Diese epistemische Arbeitsteilung funktioniert nur, wenn relevante Akteur:innen ihrer *Rolle* gerecht werden.

In Abschnitt 4.2.1 habe ich im Anschluss an Adam Carter und Emma Gordon (2011) vorgeschlagen, dass die Rolle eines Stakeholders legitime Erwartungen bezüglich seiner Expertise weckt. Eine solche Erwartung trägt mutmaßlich auch zu der von Belkoniene, Wilkenfeld und anderen angeführten Intuition bei, dass wir einer Zweitklässlerin eine falsche Überzeugung über die Entwicklung des Menschen verzeihen, über die wir bei einer Professorin der Evolutionsbiologie nicht hinwegsehen können (siehe Abschnitt 8.3.3). Gleichmaßen erwarten wir von Chefentwicklerin Lina ein umfassendes technisches Verständnis von *JobAI*, aber nur eine oberflächliche Expertise hinsichtlich der Anwendungsdomäne des Systems; bei Personalmanager Fritz ist es umgekehrt. Diese legitimen Expertiseerwartungen, so mein Vorschlag, eignen sich als limitierender Faktor hinsichtlich der für instrumentelles Verstehen zentralen Überzeugungen. Für Lina und Fritz sind demnach nicht Überzeugungen über *alle* systembezogenen Aspekte zentral, die sich in relevantem Maße auf ihr jeweiliges Desideratum auswirken, sondern nur jene, von denen wir legitimerweise erwarten können, dass Lina bzw. Fritz sie besitzen. Rollenbedingte Verstehenslücken sind demzufolge unproblematisch, weil sie durch die Expertise der anderen gewissermaßen ‚abgesichert‘ werden.

Das klappt natürlich nur, wenn tatsächlich eine funktionale epistemische Arbeitsteilung existiert. Doch können wir dies in XAI-Kontexten voraussetzen? Es scheint ja vielmehr so zu sein, dass die architektonische Opazität bestimmter KI-Systeme verhindert, dass relevante systembezogene Aspekte von Mitgliedern der epistemischen Gemeinschaft verstanden werden: Viele moderne KI-Systeme sind eine ‚Black Box‘. Entsprechend kann man sich, anders als in anderen arbeitsteilig organisierten Bereichen, diesbezüglich auch nicht auf die Expertise anderer verlassen.

An dieser Befürchtung ist durchaus etwas dran – das Black-Box-Problem motiviert nicht zuletzt die vorliegende Dissertation. Zweifelsohne werden manche KI-Systeme derzeit nicht ausreichend verstanden, um relevante Desiderata erfüllen zu können. Hier ist

Erklärbarkeit gefragt, um ein ausreichendes Maß an Transparenz sicherzustellen. Das dadurch bestenfalls gewonnene Verständnis verteilt sich dann, in Abhängigkeit von ihrer Rolle, auf die Mitglieder der epistemischen Gemeinschaft. Verfehlt Erklärbarkeit ihr Ziel, bleibt das Black-Box-Problem ungelöst – und instrumentelles Verstehen außer Reichweite. Ich bin aber zuversichtlich, dass die Opazität von KI-Systemen in vielen Fällen ausreichend reduziert werden kann, um relevante Desiderata zu erfüllen. Denn eine funktionale epistemische Arbeitsteilung in XAI-Kontexten setzt nicht voraus, dass KI-Systeme und relevante systembezogene Aspekte vollständig verstanden werden. Das Beispiel der Wissenschaft zeigt, dass (hinreichend) gesicherte Erkenntnisse mit gewissen Verstehenslücken kompatibel sind. Dasselbe dürfte auch für das Verstehen von KI-Systemen gelten.

Abschließend möchte ich hervorheben, dass die zentralen Überzeugungen eines instrumentell verstehenden Subjekts nicht zwangsläufig ‚vollkommen‘ wahr sein, sondern die Wahrheit mitunter lediglich hinreichend gut approximieren müssen (siehe Abschnitt 5.1.2). Der Grund für diese Lockerung der Faktizitätsbedingung ist, dass instrumentelles Verstehen realistischerweise für verschiedene Stakeholder erreichbar sein soll. Dafür können Verstehensermöglicher erforderlich sein, welche die tatsächliche Funktionsweise komplexer KI-Systeme vereinfacht darstellen und daher das Risiko bergen, falsche bzw. nur approximativ wahre Überzeugungen hervorzubringen.

Das lässt sich am Beispiel von Salienzkarten erläutern, die häufig für die Erklärung von Bildklassifikationen genutzt werden (etwa Abbildungen 3.2b und 3.3b). Eine solche Salienzkarte könnte z. B. den im Hintergrund sichtbaren Schnee auf einem Huskyfoto hervorheben und damit die (falsche) Klassifikation als ‚Wolf‘ erklären (Ribeiro et al., 2016, Abb. 11 auf S. 1143). Diese Erklärung ist aus verschiedenen Gründen vermutlich nur approximativ wahr: Welche Pixel die Salienzkarte in welcher Deutlichkeit hervorhebt, hängt nicht nur von dem Bildverarbeitungsprozess des Klassifikationssystems, sondern auch von der Kalibrierung der XAI-Methode ab (Bordt et al., 2022, Kap. 4.4). Dass der Schnee ausschlaggebend war, bedeutet nicht, dass die anderen Bildbereiche sich überhaupt nicht auf die Klassifikation ausgewirkt haben. Und die hervorgehobenen Bildbereiche als ‚Schnee‘ bzw. ‚Hintergrund‘ zu kategorisieren, ist lediglich eine Interpretation der menschlichen Betrachterin der Erklärung.

Dennoch kann diese Betrachterin auf Grundlage solcher lediglich approximativ wahren Überzeugungen zu der korrekten Einschätzung kommen, dass sie dieser Klassifikation nicht trauen kann. In einem solchen Fall sind die Abweichungen zwischen der tatsächlichen Funktionsweise des KI-Systems und der Erklärung bzw. der daraus resultierenden Überzeugung kontextuell vernachlässigbar. So wie das Ideale-Gas-Modell das Verhalten realer Gase unter bestimmten Bedingungen hinreichend akkurat beschreibt, können auch vereinfachende, von der Wahrheit abweichende Verstehensermöglicher korrekt genug sein, um das kontextuell erforderliche Verstehen hervorzubringen. Ich komme in Kapitel 12 auf diesen wichtigen Punkt zurück. Zunächst soll es aber um einen anderen Aspekt von Verstehen gehen: verstehensbezogene Fähigkeiten.

9 Verstehen und Fähigkeiten

Buchstäblich alle Philosoph:innen erkennen an, dass eine verstehende Person typischerweise über bestimmte Fähigkeiten (oder Fertigkeiten; siehe Fußnote 66) verfügt. Wer etwas versteht, kann meist verschiedene Schlussfolgerungen bezüglich des Verstandenen ziehen, Erklärungen formulieren oder zugehörige Probleme lösen. Viele dieser Fähigkeiten haben eine öffentliche und eine private Variante, die eng miteinander zusammenhängen: Beispielsweise kann die Fähigkeit, etwas zu erklären, einen mentalen, ‚privaten‘ Prozess meinen, oder aber das mündliche bzw. schriftliche und damit ‚öffentliche‘ Darlegen explanatorischer Zusammenhänge. Die Unterscheidung zwischen privaten und öffentlichen Fähigkeiten ist wichtig, weil sich manche Debattenbeiträge, wie nachfolgend deutlich werden wird, lediglich auf eine dieser Varianten beziehen.

Doch was sind Fähigkeiten? Zunächst einmal ist festzuhalten, dass der Fähigkeitsbegriff eng mit dem Begriff der Möglichkeit zusammenhängt. Für eine Person, die über die Fähigkeit zu φ en verfügt, ist es in einem gewissen Sinn möglich, zu φ en. Gleichzeitig muss sie nicht tatsächlich φ en, um diese Fähigkeit zu besitzen. In dieser Hinsicht ähneln Fähigkeiten Dispositionen: So wie eine Vase disponiert ist zu zerbrechen, wenn man sie fallen lässt, auch wenn sie gerade stabil auf dem Tisch steht, kann eine Person die Fähigkeit besitzen, Schlagzeug zu spielen, auch wenn sie sich gerade an einem kretischen Strand sonnt.

Zudem kann die Manifestation von Fähigkeiten, genau wie jene von Dispositionen, durch *masks* und *finks* beeinträchtigt werden (Fara, 2008; Jaster, 2020a, u. a. Kap. 2.2 bzw. 4.8): Die Fähigkeit einer Musikerin, Schlagzeug zu spielen, kann dadurch maskiert werden, dass beide ihrer Arme eingegipst sind. In diesem Fall greift etwas ‚von außen‘ in die Manifestation der Fähigkeit ein und fungiert als *mask*. In anderen Fällen führt der Versuch der Manifestation selbst dazu, dass sich die Fähigkeit nicht manifestiert. Die Fähigkeit einer Person, kopfzurechnen, manifestiert sich vielleicht nicht, weil sie aufgrund eines traumatischen Erlebnisses im Mathematikunterricht mental blockiert, sobald sie etwas ausrechnen will (die Intention, kopfzurechnen, fungiert hier als *fink*). Ansätze des so genannten neuen Dispositionalismus haben im Lichte dieser Ähnlichkeiten vorgeschlagen, Fähigkeiten durch Dispositionen zu definieren (z. B. Fara, 2008).

Doch betrachten wir zunächst klassische Analysen von Fähigkeiten. Diese werden häufig einer von zwei Gruppen zugeordnet: Die erste Gruppe umfasst Ansätze, die Fähigkeiten anhand eines subjunktiven bzw. kontrafaktischen Konditionals analysieren. Demnach verfügt ein Subjekt *S* über die Fähigkeit zu φ en genau dann, wenn *S* φ en würde, falls *S* intendiert oder versucht zu φ en. Ansätze der zweiten Gruppe analysieren Fähigkeiten mithilfe von modaler Semantik. Die Grundidee lautet hier, dass *S* über die Fähigkeit zu

φ en verfügt genau dann, wenn es eine relevante Welt bzw. eine Menge relevanter Welten gibt, in der bzw. in denen $S \varphi$ t. Neben diesen traditionellen Ansätzen wurden (neben dem bereits erwähnten neuen Dispositionalismus) in jüngerer Zeit weitere Alternativen vorgeschlagen. Beispielsweise entwickelt Romy Jaster (2020a) eine Erfolgstheorie von Fähigkeiten, derzufolge eine Fähigkeit zu ψ en darin besteht, „in einer hinreichenden Proportion der relevanten möglichen Situationen, in denen ψ zu tun ein Erfolg wäre“, zu ψ en (Jaster, 2020b, S. 445).

Ich möchte neutral bezüglich der spezifischen Natur von Fähigkeiten bleiben, da mein Vorschlag mit verschiedenen Auffassungen von Fähigkeiten kompatibel ist. Nur einen bestimmten Aspekt von Jasters Vorschlag möchte ich für hiesige Zwecke übernehmen: Während viele Autor:innen Fähigkeiten auf (intentionale) Handlungen beschränken – z. B. die Fähigkeit, einen dreifachen Axel zu springen oder Schlagzeug zu spielen –, bezieht Jaster ausdrücklich auch Nicht-Handlungsfähigkeiten ein (z. B. die Fähigkeit zu verdauen oder seinen besten Freund zu erkennen). Manche der nachfolgend diskutierten Fähigkeiten scheinen nicht oder nicht immer bewusst ausgeführte Handlungen zu sein – z. B. die Fähigkeit, relevante Begriffe oder Ausdrücke zu verstehen oder einen Schluss zu ziehen. In der Verstehensdebatte ist es aber üblich, auch solche Nicht-Handlungsfähigkeiten zu diskutieren, entsprechend stütze ich mich nachfolgend auf dieses breite Verständnis von Fähigkeiten.

Überlegungen zu verstehensbezogenen Fähigkeiten werden häufig mit der Einsicht verknüpft, dass ein charakteristisches Merkmal von Verstehen das Begreifen von Zusammenhängen ist (siehe Abschnitt 6.1.3). Erinnern wir uns noch einmal an Stephen Grimms *Volkswagen*-Beispiel, anhand dessen ich die Notwendigkeit von Begreifen für Verstehen verdeutlicht habe:

Volkswagen: „Suppose that one day I get into my 1991 Volkswagen and none of the gauges—the speedometer, tachometer, etc.—come to life; they’re all dead. I drive the car to the garage and my mechanic tells me that the reason why the gauges are dead is because I have a bad ignition switch. I then seem to have excellent reason to assent to the claim that my gauges are dead because of a bad ignition switch, even though I fail to grasp how a bad ignition switch might lead to this result. [...] As this discussion suggests, when trying to offer an account of understanding the notion of *grasping* arises almost irresistibly.“ (Grimm, 2006, S. 531–532; Hervorhebung im Original)

Prima facie scheint ein enger Zusammenhang zwischen Verstehen, Begreifen und bestimmten Fähigkeiten zu bestehen: Wenn der hypothetische Stephen Grimm aus *Volkswagen* verstünde, warum die Instrumente seines Fahrzeugs versagen, könnte er alternative, aber falsche Erklärungen ausschließen (z. B. dass sich ein Steckverbinder im Armaturenbrett gelöst hat). Er könnte erklären, wie die Instrumente mit der Zündung zusammenhängen und wie der Defekt zum besagten Totalausfall geführt hat. Vielleicht wäre er sogar dazu in der Lage, das Problem zu beheben. Angesichts solcher Intuitionen haben manche Autor:innen

vorgeschlagen, das in Verstehen involvierte Begreifen durch charakteristische Fähigkeiten zu definieren (z. B. Baumberger, 2011, 2019; Hills, 2016).

Auch meinem eigenen Vorschlag aus Abschnitt 5.1.3 zufolge stellen Fähigkeiten – neben der Verknüpftheit der geglaubten Propositionen – eine der Subdimensionen von Begreifen dar. Beide Subdimensionen hängen miteinander zusammen: Ich finde es plausibel, dass verstehensbezogene Fähigkeiten aus einem hinreichend verknüpften Überzeugungssystem erwachsen. Wenn der Protagonist aus *Volkswagen* nicht nur isolierte Überzeugungen über das Auto besitzt (dass die Instrumente ausgefallen sind, dass die Zündung defekt ist usw.), sondern wenn diese Überzeugungen darüber hinaus angemessen miteinander verknüpft sind, dann verschafft ihm dieses Überzeugungssystem gewisse Fähigkeiten – etwa die Fähigkeit, den Ausfall der Instrumente zu erklären. Fähigkeiten sind diesem Vorschlag zufolge eine Konsequenz des Überzeugungssystems eines verstehenden Subjekts.

Das restliche Kapitel gliedert sich wie folgt. In Abschnitt 9.1 argumentiere ich, dass Fähigkeiten notwendig, aber nicht hinreichend für Verstehen *simpliciter* sind. In Abschnitt 9.2 systematisiere ich verstehensbezogene Fähigkeiten anhand von sechs Clustern, die ich zusammen mit meinen Co-Autor:innen in einem interdisziplinären Aufsatz entwickelt habe (Speith et al., 2024). In Abschnitt 9.3 übertrage ich diese Überlegungen auf instrumentelles Verstehen und argumentiere, dass das Verfügen über Fähigkeiten eine dafür notwendige, wenngleich kontextsensitive Bedingung darstellt.⁹⁸

9.1 Zum Verhältnis von Verstehen und Fähigkeiten

Dass eine verstehende Person normalerweise bestimmte Fähigkeiten besitzt, ist philosophisch nicht kontrovers. Diskutiert wird lediglich, als wie eng der Zusammenhang zwischen Verstehen und verstehensbezogenen Fähigkeiten aufzufassen ist. Auf der einen Seite wurde vorgeschlagen, dass öffentliche Fähigkeiten *hinreichend* für Verstehen sind. Ich weise diese Position in Abschnitt 9.1.1 zurück. Auf der anderen Seite wird diskutiert, ob private und/oder öffentliche verstehensbezogene Fähigkeiten *notwendig* für Verstehen sind. In Abschnitt 9.1.2 plädiere ich dafür, dass das Verfügen über bestimmte Fähigkeiten in der Tat eine notwendige Bedingung für Verstehen *simpliciter* darstellt.

9.1.1 Sind Fähigkeiten hinreichend für Verstehen?

Die öffentliche Variante verstehensbezogener Fähigkeiten lässt sich extern überprüfen – man könnte den hypothetischen Stephen Grimm aus *Volkswagen* beispielsweise darum bitten, den Totalausfall der Instrumente mündlich zu erklären oder die Zündung zu reparieren. Einige Autor:innen haben im Lichte solcher Überlegungen vorgeschlagen, dass allein von außen evaluierbare Merkmale maßgeblich für Verstehenszuschreibungen sind – also

⁹⁸Vielen Dank an Astrid Schomäcker, Timo Speith, Frauke Stoll und Andreas Streinzer, deren Feedback zu den Inhalten und der Struktur dieses Kapitels äußerst hilfreich war.

dasjenige, was eine verstehende Person in der Welt tut oder tun kann (z. B. Barman et al., 2024; Delarivière & Van Kerkhove, 2021; Kuorikoski, 2012; Sleeuw, 2023; Ylikoski, 2009; Ylikoski & Kuorikoski, 2010; vgl. auch de Regt, 2015). Diesem Vorschlag zufolge ist das Verfügen über verstehensbezogene Fähigkeiten allein bereits hinreichend für Verstehen.

Ein rein behavioristischer Ansatz wird Verstehen allerdings nicht gerecht (vgl. auch Avigad, 2008, S. 330–331; van Camp, 2014, S. 99). Dies lässt sich an einigen der zuvor diskutierten Eigenschaften von Verstehen festmachen. Zum einen ist Verstehen moderat faktiv. Wir können uns Fälle vorstellen, in denen eine Person eine haarsträubend falsche Theorie über ein Phänomen glaubt und dass diese Theorie ihr aufgrund eines glücklichen Zufalls bestimmte Fähigkeiten verleiht – wir würden einer solchen Person aber sicherlich nicht zugestehen, dass sie das Phänomen versteht (ebd., Kap. 3; Wilkenfeld, 2017a, Kap. 3.2). Beispielsweise erklärt die parodistische Pseudo-Religion des Pastafarianismus die Schwerkraft durch die Existenz des fliegenden Spaghettimonsters, dessen unsichtbare Spaghettiarme alles auf die Erde drücken. Wer diese Theorie für wahr hält, kann womöglich korrekt vorhersagen, dass ein geworfener Ball zu Boden fallen wird. Die Schwerkraft versteht eine solche Person aber offensichtlich nicht.

Zum anderen erfordert Verstehen Überzeugung. Nehmen wir an, dass eine Person eine Vielzahl von validen Schlussfolgerungen über ein natürliches Phänomen ziehen kann, aber keine davon für wahr hält. Ein Beispiel wäre eine Pastafari, die Einsteins Allgemeine Relativitätstheorie versteht und daher weiß, wie diese Theorie das Phänomen der Schwerkraft erklärt. Daher könnte sie mithilfe der Theorie diverse Schlussfolgerungen bezüglich der Gravitation ziehen. Wenn die Pastafari die Allgemeine Relativitätstheorie jedoch nicht glaubt (weil sie dem Pastafarianismus anhängt), versteht sie die Theorie lediglich symbolisch. In Kapitel 7 wurde deutlich, dass eine solche Person das fragliche Phänomen (also die Gravitation) nicht versteht, weil ihr relevante Überzeugungen fehlen. Verstehen auf Fähigkeiten zu reduzieren, noch dazu auf solche, die von außen evaluierbar sind, scheidet daher aus.

9.1.2 Sind Fähigkeiten notwendig für Verstehen?

Auch wenn ich eine Reduktion von Verstehen auf Fähigkeiten nicht für sinnvoll halte, argumentiere ich in Abschnitt 9.3, dass instrumentelles Verstehen notwendigerweise private – und womöglich auch öffentliche – Fähigkeiten erfordert. Eine solche Position ist in der philosophischen Debatte zu Verstehen umstritten. Manche Philosoph:innen haben argumentiert, dass Fähigkeiten, insbesondere inferenzielle Fähigkeiten, nicht notwendig für Verstehen sind (Bourget, 2017; Kelp, 2015). In einem einflussreichen Aufsatz diskutiert Christoph Kelp ein allwissendes Subjekt, das keinerlei Schlüsse zieht, weil es bereits alles weiß. Ein solches Subjekt sei auch allverstehend, folglich seien Fähigkeiten nicht notwendig für Verstehen (Kelp, 2015, S. 3804).⁹⁹

⁹⁹ Auch Peter Kosso nutzt das Bild des allwissenden Subjekts, um Unterschiede zwischen Wissen und Verstehen zu erörtern. Er stimmt mit Kelp überein, dass ein wahrhaft allwissendes Subjekt auch Verstehen besitzt,

Dieser Einwand lässt sich mit dem hier formulierten Vorschlag leicht abwehren. Im folgenden Abschnitt werde ich eine große Zahl verstehensbezogener Fähigkeiten vorstellen. Neben diversen Fähigkeiten, die mit dem Ziehen von Schlussfolgerungen zu tun haben, zählt dazu auch die von vielen Autor:innen hervorgehobene Fähigkeit, das Verstandene zu erklären (Abschnitt 9.2.4). Es ist unplausibel, dass ein wahrhaft allverstehendes Subjekt nicht die Fähigkeit besitzt, explanatorische Zusammenhänge zumindest privat zu erläutern.¹⁰⁰

Ein Einwand von Daniel Wilkenfeld richtet sich stattdessen gegen die Notwendigkeit öffentlicher Fähigkeiten (Wilkenfeld, 2013, S. 1012). Beispielsweise könne eine von Legasthenie betroffene Person in ihrer (öffentlichen) Fähigkeit eingeschränkt sein, mit dem verstandenen Material zu arbeiten (denkbar wäre hier die Fähigkeit, Klausurfragen schriftlich zu beantworten). Ich komme in Abschnitt 9.3 auf diesen Einwand zurück.

Ein verwandter Einwand lautet, dass Fähigkeiten lediglich ein Symptom von Verstehen sind, das nicht mit dem zugrundeliegenden epistemischen Zustand verwechselt werden sollte.¹⁰¹ Ich habe nichts gegen diese Lesart einzuwenden. Wie in der Einleitung dieses Kapitels geschildert, erwachsen verstehensbezogene Fähigkeiten plausiblerweise aus einem hinreichend verknüpften Überzeugungssystem. Diesem Vorschlag zufolge sind verstehensbezogene Fähigkeiten zwar nicht konstitutiv, aber doch notwendig für Verstehen. In Abschnitt 9.3 werde ich zudem argumentieren, dass instrumentelles Verstehen im Besonderen notwendigerweise mit bestimmten, kontextabhängigen Fähigkeiten einhergeht. Denn diese Fähigkeiten sind erforderlich, damit instrumentelles Verstehen seinem Zweck, gesellschaftliche Desiderata zu erfüllen, gerecht werden kann. Das gilt auch dann, wenn sich herausstellen sollte, dass Verstehen *simpliciter* keine Fähigkeiten erfordert.

9.2 Sechs Cluster verstehensbezogener Fähigkeiten

Im Anschluss an diese übergeordneten Überlegungen möchte ich nun konkretisieren, welche Fähigkeiten eine Person auszeichnen, die versteht. In der philosophischen Literatur wird eine große Anzahl möglicher Fähigkeiten genannt. Beispielsweise listet Jeremy Avigad elf Fähigkeiten auf, über die jemand mit mathematischem Verstehen eines Beweises verfügen könnte (Avigad, 2008, S. 327–328). Vorschläge wie jener von Avigad erheben nicht den Anspruch, notwendigerweise bzw. kontextübergreifend mit Verstehen einhergehende Fähigkeiten identifiziert zu haben. Im Gegensatz dazu haben sich andere Autor:innen bemüht, Fähigkeiten auszumachen, die unabhängig vom Kontext notwendig oder zumindest charakteristisch für Verstehen sind. Ein in dieser Hinsicht einflussreicher Aufsatz stammt

weil es auch über Wissen der relevanten Verknüpfungen verfügt. Allerdings hebt Kosso im Gegensatz zu Kelp hervor, dass verstehende Subjekte sich von bloß wissenden Personen häufig dadurch unterscheiden, dass sie über spezifische Fähigkeiten verfügen (Kosso, 2002).

¹⁰⁰Vielen Dank an Eva Schmidt für eine hilfreiche Diskussion zu diesem Aspekt.

¹⁰¹Entsprechende Aussagen finden sich unter anderem in Bachmann, 2020; Hubert & Malfatti, 2023; Khalifa, 2017; Krishnan, 2020; M. P. Lynch, 2018; Riaz, 2015; Søndergaard, 2023; Stuart, 2016; Sullivan, 2018; van Camp, 2014.

von Alison Hills. Ihrem Vorschlag zufolge besitzt eine explanatorisch verstehende Person (normalerweise) sechs spezifische Fähigkeiten (Hills, 2016, S. 663). Diese Fähigkeiten umfassen das Nachvollziehen und Wiedergeben einer Erklärung sowie bestimmte inferenzielle Fähigkeiten hinsichtlich der begriffenen explanatorischen Zusammenhänge. Allerdings ist zweifelhaft, dass diese Fähigkeiten notwendig¹⁰² für jede Instanz von Verstehen sind (z. B. wenn man annimmt, dass Verstehen nicht ausschließlich durch Erklärungen erworben werden kann, siehe Abschnitt 3.3.2). Generell werden verstehensbezogene Fähigkeiten häufig im Kontext wissenschaftlichen Verstehens diskutiert. Das führt dazu, dass viele der Vorschläge Bezug auf wissenschaftliche Erklärungen, Theorien oder Modelle nehmen (z. B. de Regt, 2017; Khalifa, 2017) und dadurch nicht unmittelbar auf andere Kontexte übertragen werden können.

Auch eine dezidierte Auseinandersetzung mit verstehensbezogenen Fähigkeiten speziell für XAI-Kontexte erfolgte lange nicht. Timo Speith, Barnaby Crook, Astrid Schomäcker, Markus Langer und ich haben daher eine interdisziplinäre Literaturschau zu Verstehen und Fähigkeiten durchgeführt und auf dieser Grundlage sechs Cluster verstehensbezogener Fähigkeiten ausgearbeitet: erkennen, bewerten, vorhersagen, intervenieren, erklären und entwickeln (Speith et al., 2024, S. 5–8). Unser Vorschlag lautet, dass diese sechs Cluster jene Fähigkeiten abbilden, die für Verstehen in XAI-Kontexten relevant sind. Die Fähigkeiten der einzelnen Cluster sind dabei nicht als notwendig für jede Instanz von Verstehen aufzufassen. Vielmehr sollen sie (näherungsweise) verschiedene Verstehensgrade abbilden: Die Fähigkeiten des Erkennen-Clusters sind tendenziell weniger anspruchsvoll als jene des Entwickeln-Clusters, auch wenn der Schwierigkeitsgrad der jeweiligen Fähigkeiten je nach Kontext variieren kann (ebd., S. 8).

In diesem Abschnitt knüpfe ich an unseren Vorschlag an. Ich gebe eine Übersicht über Fähigkeiten, die von Philosoph:innen mit propositionalem, explanatorischem und/oder objektuaem Verstehen in Verbindung gebracht worden sind, und übertrage sie auf den XAI-Kontext. Dabei nutze ich eine abgewandelte Form unserer sechs Cluster, um diese Fähigkeiten zu systematisieren. Die wichtigste Abweichung gegenüber unserem ursprünglichen Vorschlag besteht darin, dass ich unser Vorhersagen-Cluster durch das Schlussfolgern-Cluster ersetze. Beim Treffen von Vorhersagen handelt es sich um eine Schlussfolgerung bezüglich künftiger Ereignisse, doch ist umgekehrt nicht jeder Schluss eine Vorhersage. Auch wenn das Treffen von Vorhersagen in XAI-Kontexten eine wichtige verstehensbezogene Fähigkeit darstellt, möchte ich an dieser Stelle ein größeres Augenmerk auf Vollständigkeit legen und zusätzliche Fähigkeiten in das Cluster integrieren.

Ich beabsichtige mit der nun folgenden Diskussion zweierlei: Zum einen möchte ich unsere Literaturschau erweitern und demonstrieren, dass die sechs Cluster hilfreich sind, um die große Anzahl von Fähigkeiten zu strukturieren, die in der philosophischen Literatur mit

¹⁰²Es ist nicht immer eindeutig, ob Hills die von ihr vorgeschlagenen Fähigkeiten als notwendig für explanatorisches Verstehen erachtet. Zumindest scheint Hills anzunehmen, dass sie notwendig für *vollständiges* explanatorisches Verstehen sind (Hills, 2016, S. 668).

Verstehen assoziiert werden. Zum anderen möchte ich einen Fundus an verstehensbezogenen Fähigkeiten zusammenstellen, auf die sich die kontextspezifische Fähigkeitsbedingung instrumentellen Verstehens beziehen kann.

Der restliche Abschnitt gliedert sich wie folgt: In Abschnitt 9.2.1 lege ich zunächst dar, warum Fähigkeiten besonders gut geeignet sind, um das in XAI-Kontexten relevante Verstehen zu charakterisieren. Anschließend betrachte ich nacheinander die sechs Cluster verstehensbezogener Fähigkeiten, die ich in leicht abgewandelter Form zusammen mit meinen Co-Autor:innen entwickelt habe: erkennen (Abschnitt 9.2.2), bewerten (Abschnitt 9.2.3), erklären (Abschnitt 9.2.4), schlussfolgern (ursprünglich ‚vorhersagen‘; Abschnitt 9.2.5), intervenieren (Abschnitt 9.2.6) und entwickeln (Abschnitt 9.2.7).

9.2.1 Vorteile verstehensbezogener Fähigkeiten für XAI-Kontexte

In unserem Aufsatz argumentieren wir, dass verstehensbezogene Fähigkeiten vier Eigenschaften aufweisen, die sie attraktiv machen, um Verstehen in XAI-Kontexten zu charakterisieren. Erstens wird der Zusammenhang zwischen Verstehen und Fähigkeiten nicht nur innerhalb der Philosophie, sondern auch in vielen anderen Disziplinen angenommen (ebd., S. 4–5). Beispielsweise wird die Fähigkeit zu erklären und das Verstandene auf neue Situationen zu übertragen in den Kognitionswissenschaften als Marker von Lernerfolg bzw. Verstehenserwerb genutzt (Rozenblit & Keil, 2002; Williams & Lombrozo, 2010). In der Erziehungspsychologie werden Fähigkeiten (z. B. bewerten oder anwenden) herangezogen, um das Verständnis von Lernenden zu überprüfen (Bloom et al., 1956; Krathwohl, 2002). Innerhalb der KI-Forschung werden Fähigkeiten angeführt, um die potenzielle Intelligenz bzw. das mögliche Verständnis solcher Systeme zu beschreiben (Pearl & Mackenzie, 2018; Thórisson et al., 2016). Und auch in der Ergonomik (die sich mit der Optimierung von Arbeitsprozessen befasst) werden Fähigkeiten (z. B. das Treffen von Vorhersagen) genutzt, um das Situationsbewusstsein von Arbeitskräften zu charakterisieren (Endsley, 1995). Das macht Fähigkeiten nicht nur für Philosoph:innen attraktiv, sondern auch für Angehörige vieler anderer Disziplinen, die in der Erklärbarkeitsforschung involviert sind.

Zweitens sind verstehensbezogene Fähigkeiten vielfältig und dadurch gut geeignet, um einen Verstehensbegriff zu entwickeln, der den in unterschiedlichen XAI-Kontexten relevanten Instanzen von Verstehen gerecht wird (Speith et al., 2024, S. 5–8). Gleichzeitig weist die Vielfalt verstehensbezogener Fähigkeiten darauf hin, dass Verstehen in dieser Hinsicht kontextsensitiv ist: Eine Systementwicklerin, die ein KI-System mithilfe von Systemdiagnose-Werkzeugen versteht, verfügt plausiblerweise über andere Fähigkeiten als eine Ärztin, welche die Klassifikation eines medizinischen KI-Systems durch eine Attributionsmethode versteht. Angesichts der großen Anzahl von kontextuell relevanten Fähigkeiten sprechen wir bewusst von ‚Clustern‘: Die sechs Cluster sind weniger als einzelne Fähigkeiten denn als Kategorien aufzufassen, unter denen ähnliche Fähigkeiten subsumiert werden können.

Der dritte Vorteil, den meine Co-Autor:innen und ich diskutieren, ist die Verbindung

zwischen verstehensbezogenen Fähigkeiten und gesellschaftlichen Desiderata (Speith et al., 2024, S. 9–10). Je nachdem, welches Desideratum in einem gegebenen Kontext erfüllt werden soll, benötigt eine verstehende Person verschiedene verstehensbezogene Fähigkeiten. Eines unserer Beispiele bezieht sich auf ein Fairness-Desideratum: Darin wird eine Job-Bewerberin von einem Empfehlungssystem als ungeeignet für die fragliche Stelle eingestuft. Wer versteht, ob diese Einstufung fair war, verfügt plausiblerweise über bestimmte Fähigkeiten – zum Beispiel die Fähigkeit, vorherzusagen, ob die Bewerberin anders eingestuft worden wäre, wenn sie ein Mann gewesen wäre. Ich komme in Abschnitt 9.3 auf den Zusammenhang zwischen Fähigkeiten und Desiderata zurück.

Viertens sind einige der öffentlichen mit Verstehen assoziierten Fähigkeiten gut geeignet, um Verstehen in XAI-Kontexten zu operationalisieren. In manchen Fällen kann es wichtig sein, zu überprüfen, ob eine Person wirklich verstanden hat – etwa, wenn viel auf dem Spiel steht oder um sicherzustellen, dass ein Verstehensermöglicher (z. B. eine XAI-Erklärung) tatsächlich Verstehen hervorgebracht hat. Dafür kann man sich nicht auf eine unzuverlässige Selbstauskunft verlassen (siehe Abschnitt 4.2.1), sondern muss Verstehen, etwa im Rahmen empirischer Studien, von außen evaluierbar machen. In unserem Aufsatz führen wir diverse Beispiele aus der XAI-Forschung an, die zeigen, dass öffentliche Fähigkeiten bereits genau zu diesem Zweck herangezogen werden und sich zur Operationalisierung eignen (ebd., S. 8–9).¹⁰³

Verstehensbezogene Fähigkeiten sind daher eine attraktive Möglichkeit, um Verstehen in XAI-Kontexten zu charakterisieren. Der restliche Abschnitt soll diese Fähigkeiten mithilfe unserer sechs Cluster konkretisieren. Dazu stelle ich nachfolgend die einzelnen Cluster sowie exemplarische zugehörige Fähigkeiten vor. Um diese Fähigkeiten plastischer zu machen und ihre Relevanz für XAI-Kontexte zu illustrieren, stütze ich mich auf die folgenden Beispiele (beim zweiten Beispiel handelt es sich um eine Abwandlung von *Bewerbung* aus Abschnitt 8.4):

Röntgen: Die Ärztin Marla verwendet ein KI-System, um auffällige Gewebestrukturen in Röntgenbildern zu identifizieren. Bei dem System handelt es sich um ein tiefes neuronales Netzwerk. Marla nutzt zusätzlich zu diesem System eine Erklärbarkeitsmethode (z. B. LIME), die Salienzkarten für die einzelnen Klassifikationen generiert. Das Röntgenbild der Lunge einer bestimmten Patientin wird vom KI-System als ‚maligne‘ klassifiziert. Die Salienzkarte hebt eine Stelle unten links im Röntgenbild hervor.

Personal: Die Informatikerin Helena arbeitet in einem Unternehmen, das ein Expertensystem für Personalabteilungen entwickelt hat. In dieses System werden Informationen aus den Bewerbungsunterlagen der jeweiligen Bewerber:innen eingespeist – zum Beispiel der Lebenslauf, Zertifikate über zusätzliche Qualifikationsmaßnahmen

¹⁰³Kristian Barman et al. (2024) machen einen ähnlichen Vorschlag: Die Autoren schlagen vor, dass wissenschaftliches Verstehen von menschlichen oder artifiziellen Agenten anhand verschiedener verstehensbezogener Fähigkeiten gemessen werden kann.

oder Arbeitszeugnisse. Anhand dieser Informationen ordnet das KI-System allen Bewerber:innen jeweils einen Wert entsprechend der vom System abgeschätzten Eignung zu. Helena verwendet verschiedene Erklärbarkeitsansätze und Werkzeuge, um das Expertensystem zu analysieren.

Diese beiden Beispiele besitzen zwei relevante Unterschiede. Erstens: In *Röntgen* geht es um explanatorisches Verstehen, während Helena aus *Personal* das Expertensystem potenziell objektual versteht. Der zweite wichtige Unterschied ist, dass es sich beim Expertensystem aus *Personal* um ein ante-hoc erklärbares System handelt, während Marla aus *Röntgen* auf eine Post-hoc-Methode angewiesen ist. Diese Unterschiede helfen dabei, verschiedene verstehensbezogene Fähigkeiten herauszuarbeiten.

Ich nutze die beiden Beispiele nicht nur, um die im Text erwähnten Fähigkeiten zu verdeutlichen. Ich greife sie auch in den nachfolgend ergänzten Tabellen auf: Da ich aus Platzgründen nicht auf alle Fähigkeiten eingehen kann, die ich in der philosophischen Literatur identifiziert habe, findet sich dort für jedes Cluster eine ausführliche Übersicht über verstehensbezogene Fähigkeiten, einschließlich Literaturhinweisen und exemplarischer Anwendung auf *Röntgen* bzw. *Personal*. Allerdings können auch die Tabellen keinen Anspruch auf Vollständigkeit erheben. Bemerkungen zu verstehensbezogenen Fähigkeiten erfolgen oft beiläufig und sind zu breit verstreut, als dass ich in diesem Rahmen eine vollständige Literaturschau leisten könnte. Dennoch hoffe ich, damit ein breites Spektrum an relevanten Fähigkeiten abzudecken.

9.2.2 Cluster 1: erkennen

Verstehen geht normalerweise mit der Fähigkeit einher, die Bedeutung eines Symbols, einer Repräsentation oder einer anderen bedeutungstragenden Einheit zu erkennen (vgl. ebd., S. 5–6). Dabei kann es sich um mündliche Rede, Text, Abbildungen, Graphen oder Ähnliches handeln. In vielen Fällen verfügt eine verstehende Person in trivialer Weise über diese Fähigkeiten – zum Beispiel, wenn sie einen Text in einer Sprache liest, der sie mächtig ist. Allerdings sind die Adressat:innen von XAI-Erklärungen mitunter mit weniger alltäglichen Repräsentationen konfrontiert, wie Salienzkarten oder numerischen Erklärungen in Form von Vektoren oder Matrizen (ebd., S. 5). Daher ist es wichtig, in XAI-Kontexten auch basale verstehensbezogene Fähigkeiten hervorzuheben. Nichtsdestotrotz sind viele Fähigkeiten des Erkennen-Clusters so grundlegend, dass sie plausiblerweise auch mit Wissen¹⁰⁴ oder

¹⁰⁴Eine explizite Bezugnahme auf Fähigkeiten findet sich auch in tugenderkenntnistheoretischer Literatur zu Wissen (siehe Sullivan, 2018, für eine Anwendung auf die Verstehensdebatte). Viele Tugendrehabilitist:innen nehmen an, dass Wissen das Resultat dispositionaler kognitiver Fähigkeiten ist – etwa zuverlässige Wahrnehmung, Erinnerung oder gültiges Schließen. Die Annahme ist also *nicht*, dass eine Person durch den Wissenserwerb neue Fähigkeiten erlangt. Bestimmte Fähigkeiten sind vielmehr eine Voraussetzung dafür, um wahre Überzeugungen in der richtigen Weise auszubilden und dadurch Wissen zu erwerben. Solche Fähigkeiten kommen damit für Verstehen als Vermögen oder Prozess in Frage (siehe Abschnitt 4.2) und sind für meine Explikation uninteressant, solange sie nicht zugleich charakteristisch für Verstehen als

Überzeugung einhergehen (Grimm, 2020, Kap. 8.6; Hills, 2016, S. 670). Eine Übersicht über zugehörige Fähigkeiten bietet Tabelle 9.1.

Fähigkeit	Beispiel (<i>Röntgen/Personal</i>)	Literatur
semantisches Verstehen	<i>Röntgen</i> : begriffliche Kompetenz hinsichtlich <i>Röntgenbild</i> , <i>maligne</i> usw.; Erfassen der Bedeutung der Salienzkarte	Bengson, 2018; Cooper, 1994, 1995; Grimm, 2020; Khalifa, 2017; Newman, 2014, 2017a,b; Stuart, 2016; Vanello, 2025; Wilkenfeld, 2013
einer Erklärung folgen	<i>Röntgen</i> : die Erklärung der XAI-Methode nachvollziehen	Baumberger, 2011; de Regt, 2017; Elgin, 2017; Gordon, 2017; Hills, 2016; Newman, 2017b; Strevens, 2013; Vanello, 2025
Repräsentationen interpretieren	<i>Personal</i> : eine programmierte Wenn-dann-Regel als Zusammenhang in der Welt interpretieren	Elgin, 2006; Malfatti, 2023
Instanziierungen erkennen	<i>Röntgen</i> : (manche) maligne Gewebestrukturen auf Röntgenbildern erkennen	Kozak, 2023; Strevens, 2024
explanatorische Zusammenhänge bzw. Abhängigkeiten identifizieren	<i>Personal</i> : das Alter der Bewerber:innen als Einflussgröße für die Beurteilung durch das System identifizieren	Grimm, 2017

Tabelle 9.1: Übersicht über Fähigkeiten des **Erkennen**-Clusters. Ergänzt sind Beispiele dafür, wie sich die jeweilige Fähigkeit im Kontext des *Röntgen*- bzw. *Personal*-Beispiels niederschlagen könnte. Die Literaturangaben verweisen auf Autor:innen, die annehmen, dass Verstehen bzw. Begreifen im Allgemeinen oder in spezifischen Kontexten mit der jeweiligen Fähigkeit bzw. einer ähnlichen Variante einhergeht oder einhergehen kann.

Kareem Khalifa argumentiert, dass Überzeugung (und damit auch Wissen und Verstehen) den ‚Besitz‘ der geglaubten Begriffe erfordert, dass man also die Bedeutung mindestens eines korrespondierenden Ausdrucks kennt und ihn richtig anwenden kann (Khalifa, 2017, S. 57; er differenziert selbst nicht zwischen Begriff und Ausdruck). Nennen wir diese Fähigkeit ‚begriffliche Kompetenz‘.¹⁰⁵ Sein Vorschlag ähnelt damit Begriffstheorien, die Begriffe als Fähigkeiten auffassen (etwa Kenny, 2010). Wer z. B. eine Erklärung versteht, kann Khalifas Vorschlag zufolge die darin enthaltenen Ausdrücke korrekt anwenden.¹⁰⁶ Catherine

epistemischen Zustand sind.

¹⁰⁵Khalifas begriffliche Kompetenz schließt auch inferenzielle Fähigkeiten ein. Wer Begriffe wie *erklären* oder *weil* ‚besitzt‘, kann z. B. auch kontrafaktische Schlüsse ziehen (Khalifa, 2017, S. 59–60). Diese Fähigkeit greife ich in meiner Diskussion des Schlussfolgern-Clusters auf (Abschnitt 9.2.5).

¹⁰⁶Diesen Gedanken formuliert bereits Ludwig Wittgenstein in Bezug auf (linguistisches) Verstehen. Eine einschlägige Übersicht über Wittgensteins Position zum Zusammenhang zwischen Verstehen und Fähigkeiten bieten Baker & Hacker, 2005, Kap. XVII.

Elgin und Federica Malfatti stellen im Kontext objektualen symbolischen Verstehens eine weitere Fähigkeit aus dem Erkennen-Cluster heraus: Wer eine Theorie versteht, kann sie interpretieren (Elgin, 2006, S. 215; Malfatti, 2023, Kap. 5.4). Malfatti zufolge erfordert das unter anderem die Fähigkeit, die Theorie auf Phänomene anzuwenden und Aussagen über den Gegenstand der Theorie abzuleiten (ebd., S. 135).

Übertragen wir diese Fähigkeiten auf unsere Beispiele. Um zu verstehen, dass das System das Röntgenbild als ‚maligne‘ klassifiziert hat, braucht Marla aus *Röntgen* begriffliche Kompetenz hinsichtlich Begriffen wie *Lunge*, *maligne* oder *Röntgenbild*. Sie muss also die (medizinische) Bedeutung korrespondierender Ausdrücke kennen und diese (in medizinischen Kontexten) korrekt anwenden können. Allerdings sind in diesem Fall nicht nur rein sprachliche Erklärungen involviert, wie sie von Khalifa berücksichtigt werden. Marla muss außerdem erkennen, was die visuellen Repräsentationen (das Röntgenbild, die Salienzkarte) bedeuten, mit denen sie konfrontiert ist. Sie muss beispielsweise erkennen, dass die von der Erklärung hervorgehobenen Bildbereiche für jene Merkmale der Eingabedaten stehen, die sich maßgeblich auf die Klassifikation als ‚maligne‘ ausgewirkt haben. In ähnlicher Form ließe sich Helenas Verstehen in *Personal* auch daran ablesen, dass sie den Programmcode des Expertensystems interpretieren und dadurch in Bezug zur systemexternen Wirklichkeit setzen kann. Wenn Helena das Expertensystem gut versteht, könnte sie beispielsweise erkennen, dass sich eine bestimmte Codezeile auf die Abschlussnote, Französischkenntnisse und den finalen Score der Bewerber:innen bezieht.

9.2.3 Cluster 2: bewerten

Verstehen geht häufig mit der Fähigkeit einher, Aspekte des Verstehensobjekts zu bewerten, also beispielsweise zu beurteilen, ob diese Aspekte angemessen oder verbesserungswürdig sind (Speith et al., 2024, S. 6). Ein Beispiel sind die Ursachen oder Gründe, die zu der Ausgabe eines KI-Systems geführt haben. Um z. B. beurteilen zu können, ob die Ausgabe korrekt ist, genügt es nicht, diese Gründe (o. ä.) zu kennen. Darüber hinaus braucht es auch relevantes Hintergrundwissen, um abschätzen zu können, ob es sich um gute oder adäquate Gründe (o. ä.) für die fragliche Ausgabe handelt. Auch die Ausgabe selbst kann durch eine verstehende Person bewertet werden. Wenn beispielsweise ein Bildklassifikationssystem ein Katzenfoto als ‚Bobtail‘ klassifiziert (siehe Abbildung 7.1), dann kann eine durchschnittliche Betrachterin diese Ausgabe als inadäquat beurteilen, auch ohne Details darüber zu kennen, was zu dieser Klassifikation geführt hat (vgl. ebd.). Tabelle 9.2 nennt weitere Fähigkeiten des Bewerten-Clusters.

In das Bewerten-Cluster fügt sich die von Khalifa diskutierte Fähigkeit ein, wissenschaftliche Erklärungen zu evaluieren (engl. *scientific explanatory evaluation*, kurz SEEing). SEEing hat drei Phasen: Zunächst ziehen Wissenschaftler:innen verschiedene mögliche Erklärungen des fraglichen Phänomens in Betracht. Anschließend vergleichen sie diese Erklärungen miteinander. Auf dieser Grundlage bilden sie schließlich passende doxastische Einstellungen aus. Khalifa ordnet SEEing dem Prozess des Wissens- bzw. Verstehens-

Fähigkeit	Beispiel (Röntgen/Personal)	Literatur
motivieren, verteidigen	<i>Personal</i> : den spezifischen Gehalt oder Aufbau des Expertensystems begründen	Avigad, 2008; Baumberger, 2011; Baumberger & Brun, 2017; Kuorikoski, 2012
als angemessen oder problematisch beurteilen	<i>Röntgen</i> : die Klassifikation als ‚maligne‘ als (in)adäquat beurteilen	Cooper, 1995; Newman, 2014; Speith et al., 2024
Relevanz einschätzen	<i>Röntgen</i> : den jeweiligen Einfluss zweier Bildbereiche auf die Klassifikation als ‚maligne‘ einschätzen	Cooper, 1994, 1995; Elgin, 2007, 2017; Moravcsik, 1979
konkurrierende Erklärungen evaluieren	<i>Röntgen</i> : die falsche Erklärung ausschließen, dass das Alter der Patientin zur Klassifikation geführt hat	de Regt, 2009; Khalifa, 2017; Newman, 2017a,b

Tabelle 9.2: Übersicht über Fähigkeiten des **Bewerten**-Clusters. Die zusätzlichen Anmerkungen bezüglich Tabelle 9.1 gelten auch hier.

werbs zu und nicht Verstehen als epistemischem Zustand (Khalifa, 2017, S. 12–13). Damit betrifft SEEing mein Explikat streng genommen nicht (siehe Abschnitt 4.2.1). Allerdings ist plausibel, dass Personen mit (hochgradigem) explanatorischem Verstehen die Fähigkeit aufweisen, konkurrierende Erklärungen zu evaluieren.

Fähigkeiten des Bewerten-Clusters könnten auch in unseren Beispielen relevant sein. Helena aus *Personal* könnte in der Lage sein, die Einstufung einer bestimmten Bewerberin durch das Expertensystem als angemessen oder unangemessen zu beurteilen, wenn sie das System und seinen Anwendungskontext gut versteht. Khalifas SEEing lässt sich am *Röntgen*-Beispiel verdeutlichen. Wenn ein Kollege von Marla die Vermutung äußert, dass das Alter der Patientin zur Klassifikation des Röntgenbilds als ‚maligne‘ geführt hat, kann Marla diese Erklärung unter Umständen ausschließen – etwa, weil sie versteht, welche Eingabedaten dem System zur Verfügung standen, oder weil keine der vom XAI-Ansatz hervorgehobenen Merkmale in irgendeinem (nicht einmal indirekten) Zusammenhang zum Alter der Patientin stehen.

9.2.4 Cluster 3: erklären

Eine für Verstehen paradigmatische Fähigkeit ist das Erklären explanatorischer Zusammenhänge (vgl. Speith et al., 2024, S. 7). Eine Evolutionsbiologin kann normalerweise die Evolutionstheorie erklären, eine typische Erkenntnistheoretikerin ist dazu in der Lage, das Gettier-Problem zu erläutern usw. Der enge Zusammenhang zwischen Verstehen und der Fähigkeit des Erklärens beruht auf der in Abschnitt 3.3.2 beschriebenen besonderen Rolle von Erklärungen für den Erwerb von Verstehen. Wenn Verstehen häufig durch Erklärungen erworben wird, ist es kaum überraschend, dass eine verstehende Person oftmals über die

9.2 Sechs Cluster verstehensbezogener Fähigkeiten

Fähigkeit	Beispiel (Röntgen/Personal)	Literatur
Erklärungen (in eigenen Worten) wiedergeben	<i>Röntgen</i> : die Erklärung der XAI-Methode in eigenen Worten wiedergeben	Baumberger, 2011, 2019; Baumberger & Brun, 2017, 2021; Carter & Gordon, 2014; Cooper, 1994; de Regt, 2015; Elgin, 2006, 2017; Gordon, 2017; Hills, 2016; Hubert & Malfatti, 2023; Khalifa, 2017; Knüsel & Baumberger, 2020; Kuorikoski, 2012; M. P. Lynch, 2018; Newman, 2017a,b; Pritchard, 2014; Stuart, 2016; Vanello, 2025; Ylikoski, 2009; Zagzebski, 2001
(erklärungs-suchende) Fragen beantworten	<i>Röntgen</i> : Kolleg:innen oder Patient:innen verschiedene Fragen zur Klassifikation beantworten	Avigad, 2008; Barman et al., 2024; Carter & Gordon, 2014; Cooper, 1994, 1995; Grimm, 2010, 2011; Kvanvig, 2003, 2009b; Moravcsik, 1979; Newman, 2014, 2017b; Ylikoski, 2009
visualisieren	<i>Personal</i> : den Aufbau des Systems schematisch darstellen	Avigad, 2008; Baumberger, 2019; Zagzebski, 2001; siehe auch de Regt, 2009, 2015, 2017; de Regt & Dieks, 2005; Zagzebski, 2019
Definitionen nennen	<i>Personal</i> : relevante Definitionen bezüglich der informatischen Grundlagen des Expertensystems nennen	Zagzebski, 2001
zusammenfassen oder ausführlicher beschreiben	<i>Personal</i> : wesentliche Aspekte des Aufbaus und der Funktionsweise des Expertensystems darstellen	Avigad, 2008; Baumberger, 2019; Cooper, 1995
generalisieren oder konkretisieren	<i>Röntgen</i> : die spezifische Erklärung für die Klassifikation als ‚maligne‘ verallgemeinern	Avigad, 2008; Cooper, 1995; de Regt & Dieks, 2005; Newman, 2014
vergleichen und kontrastieren	<i>Personal</i> : das Expertensystem mit anderen Systemarchitekturen vergleichen	Cooper, 1994, 1995; Stuart, 2016
Metaphern und Analogien formulieren	<i>Röntgen</i> : die Klassifikation des Röntgenbildes mithilfe der Netzwerk-Analogie erklären (siehe Abschnitt 2.1.2)	Cooper, 1994, 1995; de Regt & Dieks, 2005
Beispiele nennen	<i>Röntgen</i> : Beispiele für aus ähnlichen Gründen klassifizierte Röntgenbilder anführen	Avigad, 2008; Cooper, 1995; Elgin, 2017; Newman, 2014
kontextualisieren	<i>Personal</i> : das konkrete Expertensystem in der historischen Entwicklung von Expertensystemen verorten	Avigad, 2008; Kosso, 2007; Wilkenfeld, 2013; Zagzebski, 2001
systematisieren, strukturieren	<i>Personal</i> : wichtige Elemente des Expertensystems sinnvoll gruppieren	Carter & Gordon, 2014; Cooper, 1994, 1995; Moravcsik, 1979; Zagzebski, 2001, 2019

Tabelle 9.3: Übersicht über Fähigkeiten des **Erklären**-Clusters. Die zusätzlichen Anmerkungen bezüglich Tabelle 9.1 gelten auch hier.

Fähigkeit verfügt, das Verstandene zu erklären – wenigstens in dem Sinne, dass sie eine semantisch erfasste (also ‚erkannte‘) Erklärung privat oder öffentlich wiedergeben kann.

Entsprechend heben viele Autor:innen diese Fähigkeit im Kontext von Verstehen bzw. Begreifen hervor (siehe Tabelle 9.3). Interessanter scheint mir zu sein, dass in der Literatur viele weitere Fähigkeiten diskutiert werden, die sich ebenfalls als Teil von Erklärungspraktiken auffassen lassen. Beispielsweise diskutiert Neil Cooper diverse Fähigkeiten, die eng mit der Fähigkeit des Erklärens zusammenhängen – z. B. die Fähigkeiten Beispiele zu geben, Analogien zu bilden, das Verstandene zusammenzufassen oder mit anderen Dingen zu vergleichen (Cooper, 1995, S. 208–209). Hervorhebenswert ist auch die Fähigkeit der Visualisierung, die insbesondere Henk de Regt in diversen Arbeiten diskutiert hat (z. B. de Regt, 2017, Kap. 7 – er erachtet diese Fähigkeit aufgrund individuell unterschiedlicher Präferenzen aber ausdrücklich nicht als notwendig für Verstehen). Unabhängig davon, ob man bestimmte visuelle Repräsentationen selbst als Erklärung ansieht (siehe Abschnitt 3.3.2): Personen, die etwas verstehen, nutzen häufig Visualisierungen, um ihrem Gegenüber etwas zu erklären (vertraute Beispiele sind die Tafelbilder von Lehrkräften oder die Umgebungsskizzen, mit denen wir jemandem den Weg erklären).

Übertragen wir diese Fähigkeiten auf die eingangs eingeführten Beispiele. Wenn Marla aus *Röntgen* versteht, warum das KI-System das Röntgenbild der Lunge als ‚maligne‘ klassifiziert hat, könnte sie die Klassifikation (in ihren eigenen Worten) erklären – zum Beispiel durch Angabe von Gründen, Ursachen oder Mechanismen (hier kommen die unterschiedlichen Arten von Erklärungen ins Spiel, die ich in Abschnitt 3.3.1 vorgestellt habe). Die erweiterten Fähigkeiten des Erklären-Clusters lassen sich auch gut auf das *Personal*-Beispiel anwenden. Wenn Helena objektuales Verstehen des Expertensystems besitzt, könnte sie dazu in der Lage sein, Beispiele für ähnliche Systeme anzuführen, die zentralen Funktionsprinzipien des Systems zusammenzufassen, das Expertensystem mit anderen Systemarchitekturen zu vergleichen oder den Aufbau des Systems schematisch darzustellen.

9.2.5 Cluster 4: schlussfolgern

Eine verstehende Person kann typischerweise verschiedene Schlussfolgerungen hinsichtlich des Objekts des Verstehens ziehen. Jemand, der beispielsweise ein hochgradiges Verständnis eines KI-Systems hat, kann ableiten, welche Ausgaben das KI-System produzieren würde, wenn die Eingaben oder relevante Parameter anders wären. Wie eingangs erläutert, haben wir im ursprünglichen Aufsatz das Cluster auf Fähigkeiten beschränkt, die mit dem Vorhersagen des Verhaltens eines KI-Systems zusammenhängen. Um einer größeren Zahl an Fähigkeiten Rechnung zu tragen, erweitere ich das Cluster in dieser Arbeit um zusätzliche Inferenzen. Eine Übersicht bietet Tabelle 9.4.

Wohl am häufigsten genannt wird die Fähigkeit zum kontrafaktischen Schließen. Stephen Grimms Vorschlag zufolge ist diese Fähigkeit die Folge der Natur des Begreifens: Er setzt explanatorisches Verstehen mit nicht-propositionalem Wissen von Ursachen gleich

9.2 Sechs Cluster verstehensbezogener Fähigkeiten

Fähigkeit	Beispiel (Röntgen/Personal)	Literatur
durch Inferenzen verknüpfen	<i>Personal</i> : schlussfolgern, wie sich einzelne Regeln auf andere Regeln des Expertensystems auswirken	Cooper, 1994, 1995; Hills, 2016; Hubert & Malfatti, 2023; Kosso, 2002; M. P. Lynch, 2018; Malfatti, 2023; Newman, 2014, 2017a,b; Stuart, 2016; Wilkenfeld, 2013
kontrafaktisches Schließen	<i>Röntgen</i> : schließen, dass das Röntgenbild als ‚gutartig‘ klassifiziert worden wäre, wenn der auffällige Bildbereich oben rechts gewesen wäre	Barman et al., 2024; Baumberger, 2011, 2019; Baumberger & Brun, 2017; Carter & Gordon, 2014; de Regt, 2017; Egler, 2021; Elgin, 2021; Gordon, 2017; Grimm, 2006, 2010, 2011, 2014, 2017; Khalifa, 2017; Kosso, 2002, 2007; Kuorikoski, 2012; M. P. Lynch, 2018; Pérez, 2019; Riggs, 2003; Woodward, 2003; Ylikoski, 2009; Ylikoski & Kuorikoski, 2010
alternative Ereignisse ausschließen	<i>Röntgen</i> : ausschließen, dass das Röntgenbild unter aktuellen Bedingungen als ‚gutartig‘ klassifiziert wird	Egler, 2021; Grimm, 2010
das Verstandene auf neue Probleme oder konkrete Fälle anwenden, Schlüsse über ähnliche Fälle ziehen	<i>Röntgen</i> : schließen, dass ein Röntgenbild mit ähnlichen Merkmalen als ‚maligne‘ klassifiziert wird	Baumberger, 2011, 2019; Baumberger & Brun, 2017; Carter & Gordon, 2014; de Regt & Dieks, 2005; de Regt & Gijsbers, 2017; Elgin, 2007, 2017; Gordon, 2017; Grimm, 2010; Hills, 2016; Kosso, 2002, 2007; Malfatti, 2023; Moravcsik, 1979; Newman, 2014, 2017a,b; Søndergaard, 2023; Strevens, 2013; Stuart, 2016; Vanello, 2025; Wilkenfeld, 2013
(qualitative) Vorhersagen treffen	<i>Röntgen</i> : vorhersagen, dass das Röntgenbild als ‚maligne‘ klassifiziert wird	Baumberger, 2019; Baumberger & Brun, 2017, 2021; de Regt, 2009, 2015, 2017; de Regt & Dieks, 2005; de Regt & Gijsbers, 2017; Elgin, 2006; Gijsbers, 2013; Gordon, 2021; Grimm, 2017; Hubert & Malfatti, 2023; Knüsel & Baumberger, 2020; Kosso, 2007; Kuorikoski, 2012; Newman, 2017a,b; Pérez, 2019; Riggs, 2003; Stuart, 2016; Wilkenfeld, 2013; Ylikoski, 2009
Fehlerquellen identifizieren	<i>Personal</i> : herausfinden, woher ein Bug im System rührt	Riggs, 2003
Grenzen einschätzen	<i>Personal</i> : abschätzen, wie weit man die Spezifikationen des Expertensystems ausreizen kann	Baumberger, 2019; Baumberger & Brun, 2021; Elgin, 2007, 2017; Riggs, 2003; Ylikoski, 2009

Tabelle 9.4: Übersicht über Fähigkeiten des **Schlussfolgern**-Clusters. Die zusätzlichen Anmerkungen bezüglich Tabelle 9.1 gelten auch hier.

(Grimm, 2014). Explanatorisches Verstehen erfordert dann das Begreifen von modalen Zusammenhängen in der Welt. Dieses Begreifen schlägt sich in der modalen Fähigkeit nieder, Was-wäre-wenn-die-Dinge-anders-gewesen-wären-Fragen im Sinne von James Woodward zu beantworten (ebd., S. 339; zu Woodward siehe auch Abschnitt 3.3.1). Eine damit verwandte Fähigkeit besteht darin, das Verstandene auf ähnliche Fälle anzuwenden (z. B. Hills, 2016, S. 663). Natürlich setzt das ein sinnvolles Ähnlichkeitsmaß voraus: Wie Kareem Khalifa feststellt, sind auch manche oberflächlich ähnlichen Fälle offensichtlich irrelevant für eine gegebene Instanz explanatorischen Verstehens (Khalifa, 2017, S. 71–72).¹⁰⁷ Außerdem sei die Fähigkeit zur Vorhersage erwähnt, die meine Co-Autor:innen und ich in der ursprünglichen Formulierung der Cluster besonders hervorgehoben haben (Speith et al., 2024, S. 6). Beispielsweise besagt eine Formulierung von Henk de Regts Theorie der Intelligibilität, dass eine Theorie dann intelligibel ist, wenn Wissenschaftler:innen qualitative Konsequenzen der Theorie vorhersagen können, ohne exakte Berechnungen anzustellen (de Regt, 2017, S. 102).

Diese Fähigkeiten lassen sich abermals anhand unserer Beispiele verdeutlichen. Wenn die Ärztin Marla aus *Röntgen* versteht, warum das Röntgenbild als ‚maligne‘ klassifiziert wurde, könnte sie womöglich darauf schließen, ob das Bild anders klassifiziert worden wäre, wenn der auffällige Bildbereich anstatt unten links z. B. oben rechts gewesen wäre. Außerdem könnte sie die Erklärung gegebenenfalls auf weitere Röntgenbilder übertragen und dadurch auch die Klassifikation ähnlicher Bilder verstehen. Die Informatikerin Helena aus *Personal* könnte im Sinne von de Regt vielleicht qualitative Voraussagen über das Verhalten des Expertensystems treffen – zum Beispiel, wie sich das Alter der Bewerber:innen ungefähr auf die Einschätzung der Eignung auswirkt.

9.2.6 Cluster 5: intervenieren

Zumindest hohe Verstehensgrade können mit der Fähigkeit einhergehen, das Objekt des Verstehens zielgerichtet zu manipulieren, d. h. derart zu intervenieren, dass ein gewünschter Ausgang erreicht wird (Speith et al., 2024, S. 6–7). Beispielsweise könnte eine Systementwicklerin mit ausgeprägtem Verständnis ein KI-System gezielt verändern, um erwünschtes Systemverhalten zu generieren oder Bugs zu beseitigen (siehe Tabelle 9.5).

Petri Ylikoski zählt die erfolgreiche Interaktion bzw. die Kontrolle des verstandenen Phänomens zu den für explanatorisches Verstehen charakteristischen Fähigkeiten. Wer ein Phänomen versteht, kann Ylikoski zufolge gezielt intervenieren (und weiß auch um die Grenzen der Interventionsmöglichkeiten), zum Beispiel, um Probleme zu beheben. Je umfangreicher die eigene Kontrolle, desto tiefer sei das Verstehen (Ylikoski, 2009, S. 101). In ähnlicher Form geht Stephen Grimm davon aus, dass eine Person, die versteht, wie unterschiedliche Faktoren zusammenhängen, auch weiß, welche Veränderungen erforderlich sind, um einen bestimmten Zustand herbeizuführen (Grimm, 2017, S. 225). Eine

¹⁰⁷Einen ähnlichen Punkt macht Khalifa in Bezug auf kontrafaktisches Schließen (Khalifa, 2017, S. 74–75).

verwandte Fähigkeit ist Problemlösekompetenz, die viele Philosoph:innen als distinkte verstehensbezogene Fähigkeit hervorheben (siehe Tabelle 9.5). Wer Probleme lösen kann, kann das Objekt des Verstehens derart manipulieren, dass die Schwierigkeit behoben bzw. die Aufgabe gelöst wird.

Fähigkeit	Beispiel (<i>Röntgen/Personal</i>)	Literatur
kontrollieren durch gezielte Interaktion	<i>Röntgen</i> : verändern der Modellparameter, sodass das Röntgenbild als ‚gutartig‘ klassifiziert wird	de Regt, 2015; Elgin, 2021, 2022; Fleisher, 2022; Gordon, 2021; Grimm, 2006, 2011, 2017; Kuorikoski, 2012; Stuart, 2025; Wilkenfeld, 2013; Woodward, 2003; Ylikoski, 2009; Ylikoski & Kuorikoski, 2010
Fehler beheben, Probleme lösen	<i>Personal</i> : Fehler im System beseitigen	Avigad, 2008; Bachmann, 2020; Baumberger, 2019; Gordon, 2017; Kuorikoski, 2012; M. P. Lynch, 2018; Newman, 2014, 2017a,b; Søndergaard, 2023; Stuart, 2025; Wilkenfeld, 2013; Zagzebski, 2001

Tabelle 9.5: Übersicht über Fähigkeiten des **Intervenieren**-Clusters. Die zusätzlichen Anmerkungen bezüglich Tabelle 9.1 gelten auch hier.

Für unsere Informatikerin Helena aus *Personal* bedeutet das, dass sie zum Beispiel Fehler im Expertensystem beseitigen oder sonstige Veränderungen am System vornehmen könnte, um ein gewünschtes Ergebnis zu erzielen. Die Ärztin Marla aus *Röntgen* könnte, wenn sie die Fähigkeit zur Intervention besäße, beispielsweise die Architektur des KI-Systems derart manipulieren, dass das Röntgenbild als ‚gutartig‘ klassifiziert wird – vielleicht, weil die Klassifikation als ‚maligne‘ nicht korrekt ist und Marla das System verbessern möchte. (Natürlich ist es unwahrscheinlich, dass eine Ärztin über diese Fähigkeiten verfügt – dafür braucht es einen Verstehensgrad, der üblicherweise Informatiker:innen vorbehalten ist.)

9.2.7 Cluster 6: entwickeln

Ein Merkmal hoher Verstehensgrade ist schließlich die Fähigkeit, etwas von Grund auf neu zu entwickeln oder designen (Speith et al., 2024, S. 7–8). Seien es Erklärungen, wissenschaftliche Theorien und Modelle oder ein KI-System – wer eine derart komplexe Struktur kreieren kann, hat normalerweise ein tiefes Verständnis dieser Struktur (für zugehörige Fähigkeiten siehe Tabelle 9.6).

Ein Beispiel ist die Fähigkeit, eine neue Erklärung zu formulieren. Anstatt eine von anderen erarbeitete Erklärung wiederzugeben, ist hier die Fähigkeit gemeint, selbst eine Erklärung zu entwickeln. Diese Fähigkeit ist unter anderem Bestandteil von de Regts Intelligi-

Fähigkeit	Beispiel (<i>Röntgen/Personal</i>)	Literatur
eine Erklärung, Theorie oder ein Modell aufstellen bzw. entwickeln	<i>Personal</i> : ein neues Expertensystem kreieren	Barman et al., 2024; Baumberger, 2019; Cooper, 1994; de Regt, 2009, 2015, 2017; de Regt & Dieks, 2005; de Regt & Gijssbers, 2017; Knüsel & Baumberger, 2020; Kuorikoski, 2012; McCain, 2016; Strevens, 2013; Stuart, 2016; Vanello, 2025; Ylikoski, 2009; Zagzebski, 2001
neue bzw. weiterführende Fragen aufwerfen	<i>Röntgen</i> : fragen, ob die Klassifikation auf einem problematischen Proxy-Merkmal beruht	Cooper, 1994, 1995; M. P. Lynch, 2018; Stuart, 2016; siehe auch de Regt & Gijssbers, 2017

Tabelle 9.6: Übersicht über Fähigkeiten des *Entwickeln*-Clusters. Die zusätzlichen Anmerkungen bezüglich Tabelle 9.1 gelten auch hier.

bilität: Seinem Vorschlag zufolge erlaubt es eine intelligible Theorie Wissenschaftler:innen, Erklärungen zu konstruieren, die sich in die Theorie und vorhandenes Hintergrundwissen einfügen, indem sie wissenschaftliche Modelle entwickeln, die diese Theorie auf Phänomene anwenden (z. B. de Regt, 2017, Kap. 2). Auch die Fähigkeit, über das Verstandene hinauszugehen, indem man weiterführende Fragen stellt (ggf. auch ohne sie beantworten zu können), kann verstehende Personen auszeichnen (z. B. M. P. Lynch, 2018, S. 198).

Wenn wir diese Fähigkeiten auf unsere Beispiele übertragen, wird deutlich, wie anspruchsvoll die Fähigkeiten des Entwickeln-Clusters sein können. Marla aus *Röntgen* müsste dazu in der Lage sein, eigenständig eine Erklärung für die Klassifikation als ‚maligne‘ zu entwerfen, anstatt lediglich die Erklärung zu erfassen, die durch den XAI-Ansatz vorgegeben wird. Und Helena aus *Personal* müsste über die Fähigkeit verfügen, selbst ein Expertensystem von Grund auf zu programmieren. Etwas weniger fordernd erscheint die Fähigkeit, weiterführende Fragen zu stellen und dadurch das eigene Verstehen zu erweitern. Doch ohne ein gewisses Verständnis relevanter Hintergründe dürfte auch diese Fähigkeit recht eingeschränkt bleiben.¹⁰⁸

¹⁰⁸In der jüngeren, erweiterten Verstehensdebatte wurden weitere verstehensbezogene Aspekte vorgeschlagen. Für Befürworter:innen dieser Vorschläge möchte ich kurz skizzieren, wie sich diese Aspekte in eine Fähigkeitsbedingung integrieren ließen. Der erste Vorschlag ist inspiriert vom moralischen Sentimentalismus und nimmt an, dass moralisches Verstehen bzw. Begreifen auch mit affektiven und/oder motivationalen Reaktionen einhergeht (Callahan, 2018; Kauppinen, 2024; McSweeney, 2024; Vanello, 2025). Wer beispielsweise versteht, dass gewisse Praktiken rassistisch sind, wird Wut oder Trauer empfinden und unter geeigneten Umständen etwas dagegen unternehmen (Callahan, 2018, S. 452). Es ist nicht eindeutig, ob solche affektiven Reaktionen als notwendig für Verstehen gelten müssen, aber plausiblerweise können sie das Verstehen einer Person manchmal verbessern. Wenn man empathisches Einfühlungsvermögen als Fähigkeit auffasst, scheint eine Integration dieses Aspekts in die Fähigkeitsbedingung naheliegend (z. B. in das Bewerten-Cluster, insofern Fähigkeiten nach gängiger Auffassung eine evaluative Komponente

9.3 Instrumentelles Verstehen erfordert Fähigkeiten

Es ist philosophischer Konsens, dass Verstehen typischerweise mit spezifischen Fähigkeiten einhergeht. Im vorherigen Abschnitt wurde allerdings deutlich, dass die Liste an potenziell relevanten Fähigkeiten äußerst umfangreich ist. Aus philosophischer Perspektive wäre es attraktiv, wenn sich diese vielen Fähigkeiten verschiedenen Verstehensgraden zuordnen ließen, oder wenn eine Untermenge an Fähigkeiten extrahiert werden könnte, die über alle Kontexte hinweg notwendig für Verstehen sind. Ich befürchte allerdings, dass ein solches Vorhaben weitestgehend aussichtslos ist. Die möglichen Objekte des Verstehens (siehe Kapitel 6) sind ebenso heterogen wie die Kontexte, in denen wir Dinge verstehen. Die Fähigkeit der Vorhersage mag wichtig für das Verstehen von Wetterphänomenen sein, aber nicht für das Verstehen des Ersten Weltkriegs. Automechaniker:innen, die Problemlösekompetenz besitzen, verstehen den Grund für den Ausfall des Tachos auch dann, wenn sie unfähig sind, dies einem Laien zu erklären. Für das Verstehen einer Ärztin, die ihre Diagnosen gegenüber Patient:innen erläutern muss, scheint diese Fähigkeit hingegen zentral zu sein.

Die einzigen Fähigkeiten, die als kleinster gemeinsamer Nenner für kontextübergreifendes Verstehen zu taugen scheinen, sind einige basale Fähigkeiten des Erkennen-Clusters (z. B. semantisches Verstehen sprachlicher und nicht-sprachlicher Repräsentationen). Diese Fähigkeiten sind plausiblerweise auch notwendig für alle Instanzen instrumentellen Verstehens. Um beispielsweise zu verstehen, dass ein KI-System fair ist, bedarf es eines Mindestmaßes an begrifflicher Kompetenz hinsichtlich Begriffen wie *KI-System* oder *fair* bzw. *Fairness*. Um Verstehen auf Basis einer Salienzkarte zu besitzen, muss man den Inhalt dieses Verstehensermöglichers erfassen. Ein extremeres Beispiel aus dem Maschinellen Lernen mag die Idee noch besser verdeutlichen: Um zu verstehen, dass der Gradientenabstieg ein lokales Minimum einer hochdimensionalen Kostenfunktion identifiziert, genügt es nicht, dies irgendwie kognitiv zu verarbeiten (z. B. durch bloßes Lesen). Man muss auch fähig sein, die Bedeutung der enthaltenen (Fach-)Begriffe zu erkennen.

Allerdings sind solche Fähigkeiten des Erkennen-Clusters nicht die Fähigkeiten, die wir zuvorderst mit Verstehen assoziieren. Stattdessen scheinen uns je nach Verstehensobjekt und Kontext verschiedene verstehensbezogene Fähigkeiten besonders wichtig oder typisch

haben). Eine ähnliche Lesart vertritt auch Daniel Vanello, der argumentiert, dass das Verstehen moralischer Begriffe auch die Anwendung unserer affektiven und imaginativen Fähigkeiten erfordert (Vanello, 2025). Der zweite Vorschlag stammt aus der Philosophie des Geistes. Demzufolge setzt Verstehen bzw. Begreifen voraus, das Verstandene phänomenal zu erfahren (Bourget, 2017, 2024; Kauppinen, 2024; McSweeney, 2024). Phänomenales Erfahren haben manche Autor:innen wiederum durch spezifische Fähigkeiten charakterisiert, etwa David Lewis: „[...] knowing what it's like is the possession of abilities: abilities to recognize, abilities to imagine, abilities to predict one's behavior by means of imaginative experiments“ (Lewis, 1983, S. 131). Die meisten dieser Fähigkeiten fügen sich unmittelbar in unsere Cluster ein. Sollte sich also herausstellen, dass affektive Reaktionen bzw. phänomenales Erleben notwendig für Verstehen sind, scheint die Fähigkeitsbedingung diese Aspekte beherbergen zu können. Vielen Dank an Eva Schmidt und Astrid Schomäcker für hilfreiche Hinweise und Diskussionen dazu.

für Verstehen zu sein. Trotz dieser Variabilität ist das Verfügen über Fähigkeiten ein charakteristisches Merkmal von Verstehen (vgl. Delarivière & Van Kerkhove, 2021, S. 644). Ich denke daher, dass Fähigkeiten *grundsätzlich* notwendig für Verstehen sind, auch wenn wir (abgesehen von grundlegenden, verstehensunspezifischen Fähigkeiten) nicht dazu in der Lage sein sollten, *einzelne* notwendige Fähigkeiten zu identifizieren. Während ich diese Position für Verstehen *simpliciter* hier nicht verteidigen werde, ergibt sie sich für instrumentelles Verstehen natürlicherweise. Fähigkeiten, die keinen instrumentellen Wert in Hinblick auf die Erfüllung des fraglichen Desideratums haben, sind für instrumentelles Verstehen überflüssig. Der Nutzen einer Fähigkeit variiert wiederum mit dem Kontext.

In Abschnitt 8.4 habe ich vorgeschlagen, dass die Zentralität einer instrumentellen Verstehensüberzeugung vom Desideratum des Stakeholders abhängt, und dass dessen Rolle als Heuristik dienen kann, um die Menge zentraler Überzeugungen auf ein realistisches Maß einzuschränken. Diese Herangehensweise lässt sich auch auf verstehensbezogene Fähigkeiten übertragen. Betrachten wir dazu einige Beispiele.

Forschung von Judy Gichoya et al. hat ergeben, dass manche KI-Systeme anhand medizinischer Bilddaten die ethnische Zugehörigkeit der Patient:innen erkennen können (Gichoya et al., 2022). Dies birgt die Gefahr, dass solche Informationen in die Diagnosen des KI-Systems einfließen und z. B. Diskriminierung Vorschub leisten. Übertragen wir diese Problematik auf das *Röntgen*-Beispiel aus Abschnitt 9.2.1: Wir können stipulieren, dass es keinen guten Grund dafür gibt, die ethnische Zugehörigkeit der Patient:innen in die Diagnostik einzubeziehen. Die Ärztin Marla möchte nun verstehen, ob die Klassifikation des Röntgenbilds ihrer Patientin als ‚maligne‘ nichtsdestotrotz dadurch beeinflusst wurde, d. h. ob die Klassifikation in dieser Hinsicht fair ist. Verschiedene Fähigkeiten kommen dafür in Frage: Wenn Marla bewerten kann, ob die Klassifikation als ‚maligne‘ unabhängig von der Ethnie der Patientin korrekt ist; oder wenn sie erklären kann, welche Aspekte des Röntgenbilds in die Klassifikation eingeflossen sind; oder wenn Marla kontrafaktisch schließen kann, wie das Röntgenbild klassifiziert würde, wenn ihre Patientin einer anderen Ethnie angehörte – dann würde sie plausiblerweise verstehen, ob die Klassifikation des Systems in der fraglichen Hinsicht fair ist.

Die Erfüllung anderer Desiderata erfordert zumindest teilweise abweichende Fähigkeiten. Betrachten wir dazu die Informatikerin Helena aus *Personal* und nehmen an, dass sie herausgefunden hat, dass das Empfehlungssystem unfair ist und z. B. weibliche Bewerber:innen systematisch benachteiligt. Als Informatikerin ist es nun ihre Aufgabe, diesen Fehler zu beheben. Helenas instrumentelles Verstehen erfordert daher die Fähigkeit, gezielt in das System zu intervenieren und seine Fairness zu verbessern (vgl. auch Speith et al., 2024, S. 9) – zumindest sollte sie fähig sein, die erforderlichen Änderungen mental zu simulieren. Die Ärztin Marla aus *Röntgen* braucht diese Fähigkeit nicht: Ihr Desideratum ist nicht das Beseitigen von Fehlern bzw. die Verbesserung der Fairness des Systems, sondern die Überprüfung der Angemessenheit einer einzelnen Klassifikation. (Weitere Beispiele für den Zusammenhang zwischen verstehensbezogenen Fähigkeiten und der Erfüllung von Desiderata bieten wir ebd., S. 3–4, 9–10.)

An dieser Stelle ergibt sich ein ähnliches Problem wie in Abschnitt 8.4, als es um die Bestimmung der für instrumentelles Verstehen kontextuell zentralen Überzeugungen ging: Letztlich kann eine Vielzahl an verstehensbezogenen Fähigkeiten zur Erfüllung von Desiderata beitragen. Es wäre unplausibel, all diese Fähigkeiten von einem instrumentell verstehenden Stakeholder zu erwarten. Wenn die Ärztin Marla aus *Röntgen* beispielsweise selbstständig eine Erklärung für die Klassifikation des Röntgenbilds entwickeln könnte (anstatt nur die Erklärung wiederzugeben, die durch einen Verstehensermöglicher bereit gestellt wird), wäre diese Fähigkeit der Erfüllung ihres Fairness-Desideratums sicherlich zuträglich. Gleichzeitig kann diese Fähigkeit von einer Ärztin nicht verlangt werden. Es wäre daher unverhältnismäßig, diese Fähigkeit zu einer notwendigen Bedingung von Marlas instrumentellem Verstehen zu machen.

Um diese Schwierigkeit zu adressieren, können wir erneut auf die Rolle des Stakeholders und die damit verknüpften, legitimen Expertiseerwartungen zurückgreifen. Wie in Abschnitt 4.2.1 geschildert, beinhaltet die Expertise eines Stakeholders häufig spezifische Fähigkeiten. Von einer Informatikerin wie Helena aus *Personal* erwarten wir legitimerweise, dass sie fähig ist, Fehlerquellen im Empfehlungssystem zu identifizieren und normalerweise auch, dass sie gefundene Fehler beheben kann. Eine analoge Erwartung wäre in Bezug auf die Ärztin Marla aus *Röntgen* dagegen überzogen. Dafür erwarten wir von Mediziner:innen wie ihr beispielsweise, dass sie Röntgenbilder interpretieren und vermutlich auch, dass sie die Gründe für ihre (KI-gestützten) Diagnosen erklären können. Die Rolle des Stakeholders liefert zwar kein eindeutiges Kriterium, aber zumindest eine Heuristik, auf deren Grundlage sich die für instrumentelles Verstehen erforderlichen Fähigkeiten in angemessener Form einschränken lassen.

Zusammengefasst: Das Verfügen über verstehensbezogene Fähigkeiten ist eine notwendige Bedingung instrumentellen Verstehens. Welche konkreten Fähigkeiten für eine spezifische Instanz instrumentellen Verstehens erforderlich sind, hängt von kontextuellen Faktoren ab – insbesondere von dem relevanten Desideratum des Stakeholders, ggf. unter Berücksichtigung seiner Rolle.

Doch könnte man einwenden, dass verstehensbezogene Fähigkeiten zwar als übliches, nicht aber als notwendiges Kriterium für Verstehen aufgefasst werden sollten. Schließlich wurde in Abschnitt 9.1.2 deutlich, dass manche Philosoph:innen eine notwendige Fähigkeitsbedingung ablehnen. Hier sind einige an das *Personal*-Beispiel aus Abschnitt 9.2.1 anknüpfende Fälle, die auf den ersten Blick gegen eine notwendige Fähigkeitsbedingung für instrumentelles Verstehen zu sprechen scheinen:

Legasthenie: Informatikstudentin Safiya setzt sich im Rahmen ihres Studiums intensiv mit dem Empfehlungssystem für Jobkandidat:innen auseinander, das in Helenas Unternehmen entwickelt wurde. Sie analysiert den Programmcode und das Fehlerprotokoll, konsultiert Erklärungen für unterschiedliche Empfehlungen des Systems und überprüft die Wissensbasis akribisch auf Richtigkeit und Aktualität. Auf Grundlage dieser intensiven Analyse besitzt Safiya ein sehr akkurates, beinahe

vollständiges Überzeugungssystem über relevante Aspekte des Empfehlungssystems. Sie begreift genau, wie die einzelnen Komponenten des Systems miteinander interagieren und die jeweiligen Empfehlungen ausgeben. Am Ende des Semesters schreibt Safiya eine Klausur über das System. Darin soll sie die Funktionsweise des Systems erklären, kontrafaktische Aussagen über die Auswirkungen hypothetischer Veränderungen im System treffen, die Qualität der Wissensbasis beurteilen, das System in Bezug zur Geschichte von Expertensystemen im Allgemeinen setzen usw. Safiya versagt auf ganzer Linie: Sie leidet unter Legasthenie und hat daher große Schwierigkeiten, die Klausurfragen zu dem Empfehlungssystem schriftlich zu beantworten.

Krampf: Safiyas Kommilitonin Yma hat sich ebenfalls intensiv und erfolgreich mit dem Empfehlungssystem auseinander gesetzt und schreibt die gleiche Klausur. Yma will gerade den Stift ansetzen, da durchfährt ein heftiger und langwieriger Muskelkrampf ihren Arm. Der Schmerz ist so stark, dass sie keinen klaren Gedanken über das Empfehlungssystem fassen kann. Deswegen und aufgrund der eingeschränkten Beweglichkeit ihrer Schreibhand gibt Yma ein leeres Blatt ab und fällt durch die Prüfung.

Prüfungsangst: Auch Valea schreibt die Klausur über das Empfehlungssystem, und wie Safiya und Yma kennt sie sich hervorragend mit dem System aus. Wenn ihre Kommiliton:innen sie um Hilfe bei der Klausurvorbereitung bitten, kann Valea alle relevanten Aspekte des Systems erklären, beurteilen, diverse Schlüsse darüber ziehen usw. Doch Valea leidet unter extremer Prüfungsangst. Sobald ihr Verständnis überprüft werden soll, hat sie ein Blackout und kann sich an nichts mehr erinnern. So auch heute. Aufgrund ihres Blackouts gibt Valea ebenfalls ein leeres Blatt ab.¹⁰⁹

Safiya, Yma und Valea fehlen aus verschiedenen Gründen diverse Fähigkeiten, die ihr Verständnis des Empfehlungssystems signalisieren würden. Keine der drei besitzt die öffentliche Fähigkeit, Aspekte des Systems schriftlich zu erläutern, zu beurteilen usw. Yma und Valea verfügen in der Prüfungssituation noch nicht einmal über die korrespondierenden privaten Fähigkeiten – aufgrund des Blackouts bzw. der starken Schmerzen können sie auch mental keine Erklärungen formulieren oder Schlüsse über das System ziehen. Da Safiya, Yma und Valea nicht über relevante verstehensbezogene Fähigkeiten verfügen, so scheint es, fehlt ihnen (instrumentelles) Verstehen. Das ist wiederum kontraintuitiv: Alle drei verstehen das Empfehlungssystem offenkundig zu einem sehr hohen Grad. Es scheint unplausibel, jemandem aufgrund eines Muskelkrampfs, Schwierigkeiten mit Schriftsprache oder Prüfungsangst Verstehen abzusprechen.

Gehen wir also die drei Fälle durch und beginnen mit Yma, die aufgrund eines Muskelkrampfs scheinbar nicht die Fähigkeit besitzt, Aspekte des Systems zu erklären, zu bewerten

¹⁰⁹ Legasthenie ist an ein Beispiel von Daniel Wilkenfeld (2013, S. 1012) angelehnt; *Krampf* und *Prüfungsangst* sind inspiriert durch ähnliche Fälle von Romy Jaster (2020a, etwa S. 44 bzw. S. 142).

usw. Die Betonung liegt auf *scheinbar*, denn eigentlich verfügt Yma durchaus über diese Fähigkeiten, sie sind lediglich durch den Krampf *maskiert*, d. h. etwas Externes verhindert die Manifestation von Ymas verstehensbezogenen Fähigkeiten (siehe dazu die Einleitung dieses Kapitels). Da wir Yma maskierte Fähigkeiten nicht absprechen müssen, gefährdet das Vorhandensein einer solchen ‚Maske‘ auch nicht ihr instrumentelles Verstehen.

Ähnlich ist es bei Valea aus *Prüfungsangst*, wenngleich ihre Fähigkeit nicht maskiert wird. Vielmehr wirkt ihre Panik davor, zu scheitern, als *fink*: Dasjenige, das die Manifestation ihrer Fähigkeit auslöst (die Prüfungssituation), blockiert zugleich diese Manifestation (siehe dazu Jaster, 2020a, Kap. 4.8). Im Anschluss an die gängige Lesart in der philosophischen Literatur zu Fähigkeiten¹¹⁰ können wir daher auch in Bezug auf Valea sagen, dass sie die relevanten Fähigkeiten besitzt und (instrumentell) versteht, wenngleich ihre Prüfungsangst mit der Manifestation ihrer Fähigkeiten interferiert.

Bleibt noch Safiya aus *Legasthenie*. Sie verfügt eindeutig über diverse private verstehensbezogene Fähigkeiten: Sie kann im Geiste relevante Erklärungen formulieren, Vergleiche anstellen und Schlüsse ziehen. Safiya fehlen hingegen die korrespondierenden öffentlichen Fähigkeiten, ihre Überlegungen niederzuschreiben. Hier scheinen mehrere Lesarten möglich: Entweder wird die Manifestation dieser Fähigkeiten weder durch eine *mask* noch einen *fink* beeinträchtigt – Safiya verfügt schlichtweg nicht über die öffentliche Variante der genannten Fähigkeiten. Alternativ ließe sich das Fehlen öffentlicher Fähigkeiten (etwa flüssiges Lesen, Schreiben) als Maskierung der korrespondierenden privaten Fähigkeiten auffassen. So oder so kann auch *Legasthenie* angesichts der zahlreichen privaten Fähigkeiten Safiyas nicht zeigen, dass Fähigkeiten nicht notwendig für Verstehen sind.

Allerdings wirft *Legasthenie* im Kontext instrumentellen Verstehens eine interessante Frage auf: Erfordert die epistemische Erfüllung mancher gesellschaftlicher Desiderata womöglich nicht bloß private, sondern auch öffentliche verstehensbezogene Fähigkeiten? Ich möchte nicht ausschließen, dass dem so ist. Hier sind einige Desiderata, deren epistemische Erfüllung öffentliche Fähigkeiten notwendig machen könnte:

- *Informierte Einwilligung*: Ärzt:innen, die ihren Patient:innen die informierte Einwilligung in KI-gestützte Behandlungen ermöglichen müssen, benötigen womöglich die öffentliche Fähigkeit, ihren Patient:innen die Entscheidung des KI-Systems zu erklären.
- *Rechenschaft*: Entscheidungssubjekte, die durch den Fehler eines KI-Systems Nachteile erfahren haben und jemanden für diesen Fehler zur Rechenschaft ziehen möch-

¹¹⁰Die Tatsache, dass die klassische konditionale Analyse von Fähigkeiten mit *finks* und *masks* nicht umgehen kann und daher ergibt, dass Personen wie Yma und Valea nicht über die fraglichen Fähigkeiten verfügen, ist sogar einer der wichtigsten Einwände gegen diesen Ansatz (Fara, 2008; Jaster, 2020a, Kap. 2.2 und auch S. 142). Zur Erinnerung: Der konditionalen Analyse von Fähigkeiten zufolge verfügt ein Subjekt *S* über die Fähigkeit zu φ genau dann, wenn *S* φ würde, falls *S* intendiert oder versucht zu φ . Yma und Valea versuchen bzw. intendieren beide erfolglos, ihre Fähigkeiten auszuüben – entsprechend müsste die konditionale Analyse den beiden ihre Fähigkeiten absprechen.

ten, benötigen ggf. die öffentliche Fähigkeit, Unzulänglichkeiten im KI-gestützten Entscheidungsprozess gegenüber relevanten Akteur:innen (etwa Anwält:innen) zu artikulieren.

- *Evaluation:* Insofern XAI-Methoden und andere Verstehensermöglicher Verstehen hervorbringen sollen, kann es wünschenswert sein, das Verständnis der Adressat:innen dieser Ansätze zu evaluieren (siehe auch Abschnitt 9.2.1). Das setzt wiederum öffentliche Fähigkeiten voraus (z. B. die schriftliche oder mündliche Beantwortung erklärungs-suchender Fragen).

In Abschnitt 4.2.3 habe ich vorgeschlagen, dass instrumentelles Verstehen in der Regel auf die epistemische Erfüllung von Desiderata abzielt (außer in Fällen, in denen die epistemische und substanzielle Erfüllung zusammenfallen). Ich finde es nicht ganz klar, ob die oben angeführten Beispiele unabhängig von der epistemischen Erfüllung von Desiderata sind bzw. ausschließlich die substanzielle Variante betreffen und damit durch instrumentelles Verstehen nicht abgedeckt werden müssen. Denn die obigen Desiderata richten sich wenigstens indirekt auf andere Personen (Patient:innen, Anwält:innen bzw. Forscher:innen) jenseits des instrumentell verstehenden Subjekts und beinhalten daher eine kommunikative Komponente. Es könnte sein, dass die Kommunizierbarkeit des instrumentellen Verstehens zentraler Bestandteil der epistemischen Erfüllung solcher Desiderata ist. Dann wären öffentliche Fähigkeiten in diesen Fällen *per definitionem* notwendig für instrumentelles Verstehen.

Alternativ könnte man die epistemische Facette der oben aufgelisteten Desiderata anders herausgreifen: Diese bestünde jeweils darin, den fraglichen Stakeholder in die Lage zu versetzen, das Desideratum substanziell zu erfüllen. (Wir haben allerdings in Abschnitt 4.2.3 gesehen, dass diese Beschreibung der epistemischen Facette nicht in allen Kontexten plausibel ist.) Dafür genügte dann private Fähigkeiten, während lediglich die substanzielle Erfüllung öffentliche Fähigkeiten voraussetzte. Doch auch wenn öffentliche Fähigkeiten für die epistemische Erfüllung von Desiderata nicht erforderlich sein sollten, spielen sie eine besondere Rolle für die Evaluation instrumentellen Verstehens. Mutmaßlich ist die Fähigkeitsbedingung jene, die sich am besten für dessen Operationalisierung eignet (vgl. Barman et al., 2024; Speith et al., 2024). Die öffentliche Variante verstehensbezogener Fähigkeiten bietet daher einen attraktiven Ansatzpunkt, um instrumentelles Verstehen zu überprüfen – unabhängig davon, ob man diese Fähigkeiten der epistemischen oder der substanziellen Erfüllung verschiedener Desiderata zurechnet.

10 Verstehen und Rechtfertigung

Wir kommen nun zur letzten Kontroverse innerhalb der Verstehensdebatte, die ich diskutieren möchte. Sie betrifft die Frage, ob Verstehen (wie Wissen) epistemische Rechtfertigung erfordert, und wenn ja, welche Theorie der Rechtfertigung sich dafür am besten eignet.¹¹¹

Prima facie besitzt eine Person kein Verstehen, wenn ihr zugehöriges System an Überzeugungen nicht gerechtfertigt ist. Betrachten wir ein einfaches Beispiel:

Technik-Fan:¹¹² Harry ist Technik-Enthusiast und umgibt sich gern mit den neuesten KI-Gadgets. Allerdings hat er keine Ahnung von den technischen Hintergründen. Mithilfe von ChatGPT hat Harry es trotzdem geschafft, ein trainiertes KI-Modell zur Spamerkennung auf seinen Rechner zu laden und damit seine E-Mails zu klassifizieren. Eine Mail von seiner Mutter wird vom System als ‚Spam‘ klassifiziert. Harry wählt im Überschwang seiner Technik-Begeisterung spontan zwei darin enthaltene Worte aus, die seiner Meinung nach zu der Klassifikation geführt haben. Er bildet die folgende Überzeugung aus: *Die E-Mail wurde vom KI-System als ‚Spam‘ klassifiziert, weil darin die Worte ‚Gelenkschmerzen‘ und ‚Einlegesohlen‘ vorkommen.* Wie es der Zufall will, ist seine Überzeugung wahr. Außerdem kann Harry unter anderem korrekt vorhersagen, dass die Mail nicht länger als ‚Spam‘ klassifiziert werden wird, sobald er die fraglichen Worte daraus entfernt.

Harry ist offensichtlich nicht gerechtfertigt zu glauben, dass die Mail seiner Mutter aufgrund der Worte ‚Gelenkschmerzen‘ und ‚Einlegesohlen‘ als ‚Spam‘ klassifiziert wurde – schließlich hat er sich diese Erklärung einfach ausgedacht. Auch wenn er die Erklärung

¹¹¹Epistemische Rechtfertigung wird häufig ausgehend von der Frage danach diskutiert, was Wissen und wahre Überzeugung voneinander unterscheidet. Der gemeinsame Nenner zugehöriger Theorien ist die Annahme, dass Rechtfertigung dasjenige ist, was die Wahrheit einer Überzeugung wahrscheinlich macht. Davon abzugrenzen sind andere Formen der Rechtfertigung: Beispielsweise kann eine Person *pragmatisch* gerechtfertigt sein, eine Proposition zu glauben, weil es ihr nützt (z. B. ist Wunschdenken der Wahrheit nicht zuträglich, kann aber positive psychologische Effekte haben). *Moralische* Rechtfertigung kann z. B. Handlungen betreffen und diese als richtig oder falsch ausweisen. Im Folgenden geht es jedoch um die epistemische Rechtfertigung von Überzeugungen.

¹¹²Ein ähnliches Fallbeispiel entwirft Christoph Kelp als Einwand gegen Daniel Wilkenfelds (2013) Ansatz. Darin denkt sich jemand eine zufällig wahre Theorie über die Evolution des indischen Elefanten aus. Kelp zufolge versteht diese Person die Evolution des indischen Elefanten nicht (Kelp, 2015, S. 3803). Auch wenn Kelp an dieser Stelle nicht ausdrücklich Bezug auf Rechtfertigung nimmt, scheint das Problem mit der Elefantentheorie zu sein, dass es keinen guten Grund gibt, sie für wahr zu halten. Da Kelp Verstehen mithilfe von Wissen definiert (siehe Abschnitt 5.1.1), ist sein eigener Vorschlag gegen das Elefantenbeispiel gefeit.

glaubt, sie wahr ist und ihm (anscheinend) bestimmte verstehensbezogene Fähigkeiten verleiht, scheint Harry intuitiv nicht zu verstehen, warum die Mail einer bestimmten Kategorie zugewiesen wurde. Es scheint, dass Verstehen (genau wie Wissen) Rechtfertigung erfordert.¹¹³

Ich denke, dass diese intuitiv plausible Position zutreffend ist und Verstehen in der Tat Rechtfertigung erfordert. Allerdings haben dem manche Philosoph:innen widersprochen. In Abschnitt 10.1 weise ich daher zunächst Einwände zurück, denen zufolge Verstehen (im Gegensatz zu Wissen) auch ohne Rechtfertigung auskommt. Mein Fazit lautet, dass Rechtfertigung eine notwendige Bedingung für Verstehen darstellt. Wenn dem so ist, stellt sich die Frage, wie diese Rechtfertigung aufgefasst werden sollte. In Abschnitt 10.2 plädiere ich dafür, dass sich für Verstehen eine internalistische Theorie der Rechtfertigung am besten eignet, die kohärentistische und fundamentalistische Elemente kombiniert. Abschließend übertrage ich diese Überlegungen auf mein Explikat und argumentiere in Abschnitt 10.3, dass neben Verstehen *simpliciter* auch instrumentelles Verstehen Rechtfertigung erfordert.

10.1 Verstehen ohne Rechtfertigung?

Die meisten Philosoph:innen teilen die Intuition, die ich anhand von *Technik-Fan* motiviert habe: Eine Person, deren Überzeugungen nicht gerechtfertigt sind, besitzt kein Verstehen (z. B. Belkoniene, 2022, 2023; Elgin, 2017; Khalifa, 2017; Kvanvig, 2003; Park, 2017; Pritchard, 2010; Schmidt, 2026). Einzelne Autor:innen haben dieser Auffassung jedoch widersprochen und eine Reihe hypothetischer Fallbeispiele formuliert, die zeigen sollen, dass Verstehen keine Rechtfertigung erfordert. In diesem Abschnitt weise ich solche vorgeblichen Fälle von Verstehen ohne Rechtfertigung zurück. Mein Vorschlag lautet, dass diese Beispiele bestimmte Eigenschaften aufweisen, aufgrund derer sie irreführende Intuitionen wecken können. Ich werde daher versuchen, diese Intuitionen zu korrigieren.

Ich stütze mich nachfolgend also, wie in der Debatte üblich, auf die intuitive Einschätzung verschiedener Fallbeispiele. Vermutlich liefern Intuitionen aber keine abschließende Antwort auf die Frage nach der Notwendigkeit von Rechtfertigung für Verstehen, da verschiedene Philosoph:innen die relevanten Fälle teils unterschiedlich bewerten (siehe etwa die konträren Intuitionen von Hills, 2016, bzw. Hannon, 2021). Aufgrund dieser Un-

¹¹³ Gemeint ist hier doxastische Ultima-facie-Rechtfertigung (im Gegensatz zu propositionaler bzw. Prima-facie-Rechtfertigung). Diese Unterscheidungen lassen sich exemplarisch anhand von epistemischen Gründen erklären (auch wenn nicht alle Philosoph:innen annehmen, dass Rechtfertigung mit Gründen zu tun hat). *Prima-facie-Rechtfertigung* liegt vor, wenn eine Person gute Gründe für eine Überzeugung hat. Besitzt sie zusätzlich allerdings Anfechtungsgründe, ist ihre Überzeugung nicht *ultima facie* oder ‚alles in allem‘ gerechtfertigt. *Propositionale* (oder *Ex-ante*-) Rechtfertigung liegt vor, wenn eine Person gerechtfertigt ist, eine Proposition *p* zu glauben, weil sie hinreichende Gründe dafür hat – unabhängig davon, ob sie diese Überzeugung besitzt oder nicht. *Doxastische* (oder *Ex-post*-) Rechtfertigung erfordert darüber hinaus, dass eine adäquate Basierungsbeziehung zwischen Gründen und Überzeugung besteht. Um doxastisch gerechtfertigt zu sein, muss die Person tatsächlich die Überzeugung besitzen, dass *p*, und sich dabei auf diese guten Gründe stützen. Ihre Überzeugung ist dann wohlbegründet.

eindeutigkeit werde ich in Abschnitt 10.3 auch mithilfe des Zwecks von Rechtfertigung argumentieren, dass zumindest instrumentelles Verstehen notwendigerweise Rechtfertigung erfordert. Zunächst soll es aber um die intuitive Einschätzung einiger Fallbeispiele gehen. Betrachten wir dazu ein Beispiel von Finnur Dellsén, das zeigen soll, dass Rechtfertigung nicht notwendig für Verstehen ist:

Betrüger: „Bernie is a retired automobile mechanic living in a very small town in rural America. One morning Bernie reads in the local newspaper that a convicted confidence man is coming to town. The story included a picture of the man and the following warning: ‘This man will try to scam you, so don’t believe a word he says.’ The next day, the con man is driving past Bernie’s house when his car suddenly breaks down. The con man rings on Bernie’s doorbell, and Bernie opens the door. Recognizing the con man face from the newspaper, Bernie decides to stay inside his house while conversing with the con man. The con man tells Bernie what appears to be wrong with the car and solicits Bernie’s assistance. Based only on the con man’s description of the car’s behaviour immediately before the breakdown (all of which is accurate), Bernie immediately diagnoses the problem as a broken timing belt (which is correct).“ (Dellsén, 2017, S. 243)

In der Literatur werden diverse analoge Fälle diskutiert.¹¹⁴ Wir können jeweils annehmen, dass die Protagonist:innen dieser Beispielfälle (irrationalerweise) die entsprechenden Überzeugungen ausbilden, dass diese Überzeugungen (hinreichend) wahr sind und mit relevanten verstehensbezogenen Fähigkeiten einhergehen. Die übrigen Bedingungen von Verstehen wären demnach erfüllt. Gegner:innen einer notwendigen Rechtfertigungsbedingung behaupten, dass Bernie und seine Analoga Verstehen besitzen, obwohl ihnen Rechtfertigung fehlt.

Solche Fälle sind auch auf XAI-Kontexte übertragbar. In der Einleitung dieses Kapitels haben wir mit *Technik-Fan* bereits ein Beispiel für ungerechtfertigte Überzeugungen in Bezug auf KI kennen gelernt. Doch auch realistischere Fälle sind denkbar: Erinnern wir uns an das *XAI-Skepsis*-Fallbeispiel aus Abschnitt 7.1, in dem Informatikerin Cynthia an der Zuverlässigkeit ihrer eigenen Post-hoc-Methode zweifelte. Angesichts der zahllosen

¹¹⁴Dellsén formuliert zusätzlich den Fall der äußerst erfolglosen Schülerin Alice, der es zufälligerweise gelingt, den Satz des Pythagoras zu beweisen. Da sie keinen guten Grund hat, anzunehmen, dass sie richtig liegt, ist sie nicht gerechtfertigt zu glauben, dass der Beweis korrekt ist (Dellsén, 2017, S. 242). Analog dazu sind auch die bereits in Fußnote 74 erwähnten Beispiele: zum einen der Fall des unfähigen Kriminalkommissars Klaus, der durch bloßes Glück erstmals den wahren Mörder identifiziert, aber nicht gerechtfertigt ist, seine (wahre) Hypothese zu glauben (Malfatti, 2023, S. 190–191); zum anderen jener der Ärztin, die einen Anfechtungsgrund für ihre Überzeugung besitzt, dass einer ihrer Patienten unter einem Eisenmangel leidet (ebd., S. 188–190). Und Alison Hills diskutiert den Fall einer Schülerin, die das irreführende Zeugnis ihrer (grundsätzlich vertrauenswürdigen) Lehrerin ignoriert und entsprechend einen unangefochtenen Anfechtungsgrund für ihre (wahre) Überzeugung besitzt (Hills, 2016, S. 672). Die relevante Gemeinsamkeit dieser Beispiele scheint mir zu sein, dass die Überzeugungen der jeweiligen Protagonist:innen *ultima facie* nicht gerechtfertigt sind; entsprechend beschränke ich mich hier auf das *Betrüger*-Beispiel.

Schwächen existierender Erklärbarkeitsansätze sind solche Bedenken durchaus berechtigt (siehe Abschnitt 3.2.1). Man kann daher zumindest in Frage stellen, ob bestimmte Post-hoc-Methoden geeignet sind, um Überzeugungen über die Entscheidungsprozesse von architektonisch opaken KI-Systemen hinreichend zu rechtfertigen. Wenn Gegner:innen einer notwendigen Rechtfertigungsbedingung für Verstehen Recht haben, könnten Stakeholder in XAI-Kontexten auch dann Verstehen besitzen, wenn sie gute Gründe haben, die Akkuratheit der herangezogenen XAI-Erklärungen anzuzweifeln.

Ich denke nicht, dass diese Position einer genaueren Betrachtung standhält. Daher möchte ich zwei Erklärungsansätze vorstellen, die erklären können, weswegen Beispiele wie *Betrüger* die irreführende Intuition wecken können, dass Verstehen keine Rechtfertigung erfordert. Diesen Erklärungsansätzen zufolge erschwert es die Formulierung solcher Fallbeispiele, verschiedene Objekte bzw. Aspekte von Verstehen auseinanderzuhalten. Trennt man diese deutlicher voneinander, ist es weit weniger plausibel, dass Verstehen auch ohne Rechtfertigung auskommt.

Den ersten Erklärungsansatz übernehme ich von Miloud Belkoniene, der auch Dellséns *Betrüger*-Beispiel diskutiert. Dellsén bietet folgende Analyse des Falls an: Bernie besitzt Verstehen, weil die Evidenzen, die Rechtfertigung verhindern, explanatorisch irrelevant für das Objekt des Verstehens sind – die fehlende Vertrauenswürdigkeit des Betrügers hat nichts mit dem Defekt seines Autos zu tun. Bernie begreift die erforderlichen explanatorischen Zusammenhänge, egal ob seine Überzeugungen gerechtfertigt sind oder nicht. Aus diesem Grund, so Dellsén, gestehen wir Bernie auch in Abwesenheit von Rechtfertigung Verstehen zu (Dellsén, 2017, S. 244–245).

Dellséns Analyse bildet den Ausgangspunkt für Belkonienes Kritik. Seinem Vorschlag zufolge verwischt *Betrüger* die Grenze zwischen zwei notwendigen Bedingungen von Verstehen: erstens, dass Bernie den (tatsächlichen) explanatorischen Zusammenhang begreift und zweitens, dass er gerechtfertigt ist, den Defekt des Autos überhaupt für erklärungsbedürftig zu halten (Belkoniene, 2022, S. 5).¹¹⁵ Die Vermischung dieser beiden Aspekte löse die irreführende Intuition aus, dass Bernie auch in Abwesenheit von Rechtfertigung versteht. Belkoniene entwirft daher alternative Fälle, in denen der Unterschied zwischen Begreifen und Rechtfertigung deutlicher hervortritt. Hier ist eines seiner Beispiele, das er an einen Fall von Stephen Grimm (2006, S. 520) anlehnt:

Spezialkaffee: „Sam has been drugged and has undergone various hallucinations since he drank the coffee which contained the drug. This has led him to the conclusion that one of his friends drugged his coffee and, after calling the friend at issue, he received confirmation that his coffee was drugged. Still under the influence of the drug, Sam hallucinates a glass falling on the floor of his kitchen and breaking. [...] [B]y sheer luck, a glass really fell on his kitchen floor and broke. Sam, who

¹¹⁵Belkoniene kritisiert zudem weitere Argumente von Dellsén, die sich auf deduktive Stichhaltigkeit (engl. *deductive cogency*) stützen (z. B. Dellsén, 2018). Aus Platzgründen lasse ich diesen Argumentationsstrang hier außen vor und verweise auf Belkonienes (2022, Kap. 3) Einwände.

cannot refrain from believing the content of his hallucinatory experience, forms the belief that a glass fell on his kitchen floor and broke and endorses a correct account of why the glass broke.“ (Belkoniene, 2022, S. 4)

In Fällen wie *Spezialkaffee* scheint es weit weniger plausibel, dass Verstehen ohne Rechtfertigung auskommt. Belkoniene zufolge hat das den folgenden Grund: *Spezialkaffee* verdeutlicht, dass es unplausibel ist, einer Person Verstehen-warum p zuzusprechen, wenn sie nicht einmal gerechtfertigt ist zu glauben, dass p (einen ähnlichen Punkt macht Park, 2017, S. 381). Sam ist nicht gerechtfertigt zu glauben, dass ein Glas auf seinem Küchenboden zersprungen ist – entsprechend versteht er auch nicht, warum das Glas zerbrochen ist, obwohl er den tatsächlichen explanatorischen Zusammenhang (zufälligerweise) begreift (Belkoniene, 2022, S. 3–5). Wenn man Begreifen und Rechtfertigung besser voneinander trennt, wird deutlich, dass Verstehen intuitiv beide Aspekte erfordert.

Hier könnte man einwenden, dass Belkonienes Argument in Bezug auf explanatorisches Verstehen in XAI-Kontexten nicht immer greift. Es besteht nämlich ein interessanter Unterschied zu Fällen wie *Betrüger* oder *Spezialkaffee*: In XAI-Kontexten ist man typischerweise gerechtfertigt, das Explanandum (p) zu glauben, weil es sich dabei häufig um die Ausgabe (z. B. eine Klassifikation) eines KI-Systems handelt. Anders als Bernie oder Sam sind Stakeholder in XAI-Kontexten also üblicherweise gerechtfertigt, das Explanandum für erklärungsbedürftig zu halten. Angezweifelt wird vor allem, ob die von XAI-Methoden generierten Erklärungen akkurat sind – d. h. ob diese Erklärungen Überzeugungen über das Explanans (q) rechtfertigen können. Demnach betrifft die potenziell unzureichende Rechtfertigung, anders als in den klassischerweise in der Verstehensdebatte diskutierten Fällen, häufig nur eine Teilmenge der putativen Verstehensüberzeugungen – nämlich jene über das Explanans. *Technik-Fan* legt nahe, dass Verstehen auch in solchen Fällen Rechtfertigung erfordert. Doch gibt es eine andere, auch auf gängige XAI-Kontexte anwendbare Möglichkeit, um die (irreführenden) Intuitionen hinsichtlich Fällen wie *Betrüger* zu erklären?

Das bringt mich zum zweiten Erklärungsansatz, der (anders als der erstgenannte) von mir stammt. Er stützt sich auf die bereits ausführlich behandelte Unterscheidung zwischen faktischem und symbolischem Verstehen (siehe Abschnitt 6.2 und Kapitel 7 und 8). Wir sind demnach geneigt, Bernie aus *Betrüger* Verstehen zuzuschreiben, weil er in gewisser Hinsicht Verstehen besitzt – nämlich *symbolisches* Verstehen. Bernie begreift die Zusammenhänge zwischen dem Liegenbleiben eines Autos und einem defekten Zahnriemen. Das erlaubt ihm, symbolisch eine ‚Theorie‘ oder mögliche Erklärung bezüglich des (hypothetischen) Defekts des Betrüger-Autos zu verstehen.

Für sich genommen – also unabhängig von ihrer konkreten Anwendung auf das Auto des Betrügers – besitzt diese Theorie einen hohen Grad an Rechtfertigung: Wir können annehmen, dass Bernies Überzeugungen darüber, was mit einem Auto passiert, wenn der Zahnriemen reißt, dank seiner Ausbildung und Berufserfahrung sehr gut gerechtfertigt sind. Allerdings ist Bernie nicht gerechtfertigt, diese Theorie als akkurate Repräsentation

dessen aufzufassen, was mit dem Auto des Betrügers geschehen ist. Bernie versteht daher nicht *faktisch*, warum das Auto liegen geblieben ist. Hier sind zwei Varianten (V1 und V2) von *Betrüger*, die es erleichtern, symbolisches und faktisches Verstehen auseinander zu halten:

Betrüger (V1): Automechanikerin Zoe ist eine ehemalige Kollegin von Bernie. Bernie erzählt ihr von seinem Erlebnis mit dem Betrüger. Er erklärt ihr ausführlich seine Theorie, derzufolge der defekte Zahnriemen aufgrund von Verschleiß gerissen ist, in der Folge die Ventile nicht mehr gesteuert wurden und dadurch der Motor kaputt gegangen ist. Da Zoe eine kompetente Automechanikerin ist, versteht sie Bernies Theorie mithilfe seiner Erläuterungen ebenso gut wie Bernie selbst. Doch Bernie erzählt ihr auch, dass er diese Theorie anhand der Aussagen eines bekannten Betrügers formuliert hat und diese Aussagen sehr wahrscheinlich gelogen waren. Zoe ist sich daher sicher, dass das Auto des Betrügers nicht aufgrund eines gerissenen Zahnriemens liegen geblieben ist.

Betrüger (V2): Dank Bernies Erläuterungen versteht Zoe dessen Theorie über den Motorschaden des Autos, und sie weiß, dass Bernie diese Theorie auf Grundlage der Aussagen eines notorisch lügenden Betrügers aufgestellt hat. Allerdings ist Zoe eine große Auto-Enthusiastin und sperrt sich innerlich gegen die Vorstellung, dass jemand in Bezug auf ihr absolutes Lieblingsthema – Viertaktmotoren – lügen könnte. Sie redet sich daher ein, dass der Betrüger dieses eine Mal die Wahrheit gesagt hat und Bernies Theorie stimmt. Zoe versteht nicht bloß Bernies Theorie, sondern glaubt auch, dass sie wahr ist und der gerissene Zahnriemen einen Motorschaden und das Liegenbleiben des Autos verursacht hat. Wie es der Zufall will, ist diese Theorie tatsächlich wahr, auch wenn Zoe davon nichts ahnt.

Beide Varianten scheinen mir klare Fälle symbolischen Verstehens zu sein. Zoe aus V1 versteht sehr gut, warum Autos aufgrund eines gerissenen Zahnriemens liegen bleiben. Sie versteht, wie Verschleiß, Zahnriemen, Ventilsteuerung und Motorschaden grundsätzlich zusammenhängen. Sie versteht daher, warum das Auto des Betrügers liegen geblieben *wäre*, wenn es das vom Betrüger geschilderte Verhalten gezeigt *hätte*. Sie hat aber keinen guten Grund anzunehmen, dass dem so war. In V2 besitzt Zoe dasselbe symbolische Verstehen, doch diesmal hält sie Bernies Theorie (irrationalerweise) für wahr. Trotzdem scheint Zoe aus V2 (genau wie in V1) intuitiv kein faktisches Verstehen des Motorschadens zu besitzen.

Der Vergleich der beiden Varianten von *Betrüger* legt nahe, dass eine unzureichende Trennung von symbolischem und faktischem Verstehen manche irreführende Intuitionen hinsichtlich bestimmter Fallbeispiele erklären kann. Vielleicht erfasst diese Unterscheidung auch das Unbehagen, das Seungbae Park bezüglich Dellséns Position geäußert hat: Wenn man Verstehen unabhängig von Rechtfertigung mache, so Park, erlaube Verstehen keinerlei Schlussfolgerungen über oder Handlungen in der Welt. Verstehen sei dann losgelöst von der Wirklichkeit (Park, 2017, Kap. 2). Dem zweiten Erklärungsansatz zufolge ergibt sich

diese Entkopplung von der Welt dadurch, dass es sich lediglich um Fälle symbolischen Verstehens handelt.

Hervorhebenswert ist, dass der zweite Erklärungsansatz auch greift, wenn die Wahrheit des Explanandums außer Frage steht. Nehmen wir eine Person wie Cynthia aus *XAI-Skepsis*, die berechtigte Zweifel gegenüber Post-hoc-Methoden hegt. Solche Stakeholder sind in der Regel gerechtfertigt zu glauben, dass das zu erklärende KI-System eine bestimmte Ausgabe produziert hat. Außerdem sind sie normalerweise gerechtfertigt zu glauben, dass *laut der XAI-Methode* bestimmte Aspekte der Eingabedaten ausschlaggebend für die Klassifikation waren (beispielsweise, weil sie eine Salienzkarte betrachten). Wir können solchen Stakeholdern symbolisches Verstehen einer (möglichen) Erklärung zugestehen, ohne ihnen zugleich faktisches Verstehen der Entscheidungsprozesse des fraglichen KI-Systems zuschreiben zu müssen. Sind die Stakeholder hingegen gerechtfertigt zu glauben, dass die fragliche XAI-Erklärung wahr ist, eröffnet dies die Möglichkeit, die Klassifikation auch faktisch zu verstehen.

Fassen wir noch einmal zusammen: Die obigen Überlegungen sprechen dafür, dass Verstehen – *pace* Dellsén und Co. – durchaus Rechtfertigung erfordert. Der Grund für die gegenläufigen Intuitionen bezüglich mancher Fallbeispiele ist, dass man eine hypothetische Erklärung oder Theorie begreifen (der erste Erklärungsansatz) und symbolisch verstehen kann, ohne damit faktisch einen Aspekt der Wirklichkeit zu verstehen (der zweite Erklärungsansatz).

10.2 Welche Rechtfertigungstheorie eignet sich für Verstehen?

Wenn die Argumentation des vorherigen Abschnitts stimmig ist, stellt sich die Frage, welche Theorie der Rechtfertigung für Verstehen geeignet ist. In diesem Abschnitt schlage ich eine in meinen Augen attraktive Position vor: Verstehen erfordert eine internalistische, hybride Form der Rechtfertigung, die fundamentalistische und kohärentistische Elemente kombiniert. Um diesen Vorschlag einordnen zu können, bedarf es zunächst einer kurzen Einführung in relevante Grundsatzdebatten zur Rechtfertigung von Überzeugungen (Abschnitt 10.2.1). Anschließend gehe ich in Abschnitt 10.2.2 auf den Zusammenhang zwischen dem für Verstehen charakteristischen Begreifen und der Rechtfertigung von Verstehensüberzeugungen ein. Schließlich folgt in Abschnitt 10.2.3 mein Plädoyer für einen internalistischen, hybriden Ansatz.

10.2.1 Grundfragen der Theorie der Rechtfertigung

Innerhalb der philosophischen Forschung zu Rechtfertigung existieren zwei Grundsatzdebatten, die für das Verständnis meines Vorschlags relevant sind: Ist Rechtfertigung internalistisch oder externalistisch? Und, sofern Rechtfertigung internalistisch ist, wie ist

sie strukturiert?

Internalismus versus Externalismus

Die Unterscheidung zwischen Internalismus und Externalismus betrifft die Frage danach, durch welche Entitäten die Rechtfertigung von Überzeugungen zustande kommt. *Internalistische* Theorien der Rechtfertigung gehen davon aus, dass interne Faktoren der fraglichen Person eine doxastische Einstellung rechtfertigen. Die meisten Internalist:innen vertreten eine Variante des *Evidenzialismus* (z. B. Conee & Feldman, 2004). Dessen Grundidee lautet, dass die Rechtfertigung einer Überzeugung von der Gesamtmenge der Belege, die das epistemische Subjekt zum fraglichen Zeitpunkt besitzt, abhängt.

Eine wichtige internalistische Position ist der *Mentalismus* (Conee & Feldman, 2001), demzufolge Rechtfertigung von mentalen Zuständen, zum Beispiel Überzeugungen oder Sinneswahrnehmungen, abhängt. Anspruchsvoller ist der *Zugangsinternalismus* (z. B. Bonjour, 1985). Diese Position besagt, dass die rechtfertigenden Faktoren dem epistemischen Subjekt z. B. reflexiv zugänglich sein müssen.

Anders als internalistische Ansätze nehmen *externalistische* Theorien an, dass Rechtfertigung nicht oder nicht allein auf internen Faktoren beruht, sondern (auch) auf dem Verhältnis zwischen mentalen Zuständen und Aspekten der externen Welt. Externalist:innen zufolge betrifft Rechtfertigung vornehmlich den Ursprung der fraglichen Überzeugung oder die Mechanismen, welche die Überzeugung herbeigeführt haben. Die Faktoren, die relevant für die Rechtfertigung einer Überzeugung sind, müssen dem epistemischen Subjekt also nicht kognitiv zugänglich sein, und eine Veränderung der Rechtfertigung kann auch aus Veränderungen in nicht-mentalenen Faktoren resultieren.

Eine einflussreiche externalistische Theorie ist Alvin Goldmans *Prozessreliabilismus* (z. B. Goldman, 1979). Goldmans Vorschlag besteht darin, Arten von Prozessen der Überzeugungsbildung zu identifizieren, die im Großen und Ganzen zuverlässig sind – z. B. Wahrnehmung oder Erinnerung. Zuverlässigkeit meint hier, dass solche Prozesse wahrscheinlich (wenngleich nicht immer) in wahren Überzeugungen resultieren und daher rechtfertigende Kraft besitzen.

Fundamentalismus versus Kohärentismus

Die zweite hier relevante Grundsatzdebatte ist nur für internalistische Ansätze von Bedeutung. Sie betrifft die grundlegende Struktur der Rechtfertigung. Die beiden Standardpositionen sind der Fundamentalismus und der Kohärentismus, von denen wiederum verschiedene Varianten existieren. Ich möchte hier nur kurz deren übliche Ausgestaltung skizzieren. Dem *Fundamentalismus* zufolge existieren basale gerechtfertigte Überzeugungen, welche die Grundlage für die Rechtfertigung aller weiteren Überzeugungen bilden. Basale Überzeugungen sind nicht durch (wiederum zu rechtfertigende) Überzeugungen gerechtfertigt, sondern z. B. durch Wahrnehmungen, die selbst keiner Rechtfertigung bedürfen.

Der *Kohärentismus* lehnt in seiner gängigsten (d. h. holistischen) Form das Bild einer solchen linearen Begründungskette ab. Rechtfertigung ähnele vielmehr einem in sich kohärenten Netz. Die Kohärenz von Überzeugungen erfordert einer gängigen Auffassung nach Widerspruchsfreiheit sowie eine adäquate Verknüpfung zwischen den involvierten Überzeugungen. Letzteres meint üblicherweise, dass die Überzeugungen inferenziell bzw. probabilistisch sowie explanatorisch durch andere Überzeugungen des Systems gestützt werden (BonJour, 1985, S. 95–101). Einzelne Überzeugungen sind dann durch ihre Position im Netzwerk gerechtfertigt. Anders als beim Fundamentalismus kommen dadurch für viele Kohärentist:innen ausschließlich andere Überzeugungen als rechtfertigender Grund in Frage, nicht aber sonstige mentale Zustände (z. B. Wahrnehmungen).

Die meisten Fundamentalist:innen teilen auch kohärentistische Annahmen (ebd., S. 94); ein Punkt, der später noch wichtig wird. Beispielsweise charakterisiert Roderick Chisholm die fundamentalistische Kernthese wie folgt: „The knowledge which a person has at any time is a structure or edifice, many parts and stages of which help to support each other, but which as a whole is supported by its own foundation“ (Chisholm, 2008, S. 80). Fundamentalist:innen akzeptieren üblicherweise, dass Kohärenz dazu beitragen kann, Überzeugungen zu rechtfertigen – es braucht aber ein Fundament, das unabhängig davon gerechtfertigt ist, damit Kohärenz ihre Wirkung entfalten kann (Van Cleve, 2014, S. 256).

Laurence BonJour unterscheidet drei Varianten des Fundamentalismus, die Kohärenz jeweils unterschiedlichen Stellenwert einräumen (BonJour, 1985, Kap. 2.2): Der *starke Fundamentalismus* nimmt an, dass basale Überzeugungen unfehlbar, unbezweifelbar oder unkorrigierbar sind. Der *moderate Fundamentalismus* besagt, dass basale Überzeugungen zwar anfechtbar sind, grundsätzlich aber hinreichend gerechtfertigt sein können, um z. B. Wissen zu ermöglichen. Der *schwache Fundamentalismus* behauptet, dass basale Überzeugungen für sich genommen nur schwach gerechtfertigt sind. Es handelt sich lediglich um ‚Ausgangsüberzeugungen‘ (engl. *initially tenable beliefs*), die erst in Kombination mit Kohärenz die für Rechtfertigung nötige Schwelle überschreiten können.

10.2.2 Begreifen und Rechtfertigung

Die einschlägigen Theorien der Rechtfertigung haben allesamt Wissen im Blick. Sofern man sich nicht auf die kontroverse Position festlegen möchte, dass Verstehen eine Form des Wissens ist (mehr dazu in Abschnitt 11.1.3), stellt sich die Frage, welche dieser Theorien für Verstehen geeignet ist. Ein Blick in die Verstehensliteratur legt folgende Antwort nahe: alle. Es gibt bereits viele verschiedene Vorschläge, die evidenzialistische (Baumberger, 2019; Belkoniene, 2023; Newman, 2017a; Schmidt, 2026), kohärentistische (Baumberger, 2019; Kvanvig, 2003; Schmidt, 2026), reliabilistische¹¹⁶ (Baumberger, 2019; Grimm, 2017) und/oder

¹¹⁶Juan Durán und seine Co-Autor:innen haben zudem bestimmte reliabilistische Annahmen in Form des *computationalen Reliabilismus* auf KI-Systeme übertragen (Durán & Formanek, 2018; Durán & Jongsma, 2021). Die Grundidee lautet, dass wir gerechtfertigt sind, den Ausgaben eines KI-Systems zu vertrauen, wenn der Algorithmus meistens vertrauenswürdige Ergebnisse liefert (Durán & Jongsma, 2021, S. 332; vgl.

tugenderkenntnistheoretische (Khalifa, 2017; Kvanvig, 2009b; Pritchard, 2014; Riggs, 2003; Zagzebski, 2001) Elemente anführen, um die Rechtfertigung von Verstehensüberzeugungen theoretisch zu fassen.

Diese Ansätze erben jeweils die bekannten Stärken und Schwächen der zugrunde gelegten Rechtfertigungstheorien. Anstatt das bereits gut dokumentierte Für und Wider dieser Vorschläge zu diskutieren (einen Überblick bieten z. B. Feldman, 2003; Grundmann, 2017; McCain, 2016; Pappas, 2023), möchte ich einen positiven Vorschlag machen: Das in Verstehen involvierte Begreifen bietet aus meiner Sicht gute Gründe, internalistische – konkret: kohärentistische – Ansätze gegenüber konkurrierenden Vorschlägen vorzuziehen.

Verstehen hat ein internalistisches Element

Ich denke, dass sich internalistische Ansätze besser für Verstehen eignen als externalistische Theorien wie z. B. der Reliabilismus. Letztere Ansätze konzentrieren sich auf das Verhältnis zwischen Überzeugungen und Außenwelt. Intuitiv hat Verstehen nach Ansicht vieler Philosoph:innen aber ein dezidiert internalistisches Element (Baumberger & Brun, 2017; Kvanvig, 2003; Pritchard, 2010; Zagzebski, 2001; für eine Kritik siehe Grimm, 2017). Ich denke, Kvanvig hat Recht damit, dass diese Intuition von dem in Abschnitt 6.1.3 geschilderten Begreifen von Zusammenhängen herrührt, das für Verstehen derart charakteristisch ist:

What is distinctive about understanding, once we have satisfied the truth requirement, is internal to cognition. It is the internal seeing or appreciating of explanatory and other coherence-inducing relationships in a body of information that is crucial for understanding. (Kvanvig, 2003, S. 198)

[...] there is a psychological requirement concerning the coherence relations involved in understanding, to the effect that the person in question must grasp them. The way in which all the information fits together must be part of what the person is aware of. (ebd., S. 202)

Kvanvig zufolge ist ein wichtiger Grund für diese internalistischen Intuitionen der holistische Charakter von Verstehen. Wenn wir etwas verstehen, glauben wir nicht bloß atomistische Propositionen, sondern besitzen ein Netzwerk an Überzeugungen, das eine gewisse Komplexität und Verflechtung aufweist. Ein solches Netzwerk ist etwas dem epis-

auch Grote et al., 2024, S. 5–6). Dafür gebe es mindestens vier Zuverlässigkeitsindikatoren: (1) Verifikation und Validierung, also die Überprüfung des Systems durch formale Methoden und Abgleich mit etablierten Benchmarks; (2) Robustheits-Analyse, d. h. die Messung der Robustheit des Modells unter einer Vielzahl an Umständen; (3) vergangene erfolgreiche Implementierungen, um sicherzustellen, dass das System in Einklang mit etablierten Methodologien und Praktiken der relevanten Domäne entwickelt und verwendet wird; und (4) Übereinstimmung mit Expertenwissen, also dass Expert:innen dem System angemessen vertrauen (Durán & Jongsma, 2021, S. 332; Grote et al., 2024, S. 6; vgl. auch Krishnan, 2020, S. 496, 498, für ähnliche Überlegungen).

temischen Subjekt Internes.¹¹⁷ Es erscheint zudem plausibel, dass uns diese Verknüpfungen (wenigstens in Teilen) epistemisch zugänglich sind – wir *begreifen* die Zusammenhänge, und verfügen deswegen über charakteristische Fähigkeiten (siehe Kapitel 9). Externalistische Ansätze scheinen diesem Charakteristikum von Verstehen nicht gerecht zu werden.

Verstehen hat ein kohärentistisches Element

Der holistische Charakter von Verstehen überlappt stark mit Annahmen, die aus kohärentistischen Theorien der Rechtfertigung bekannt sind (Hannon, 2021, S. 280–281; Kvanvig, 2003, S. 192). In Abschnitt 6.1.3 wurde deutlich, dass die Objekte von Verstehen nicht atomistisch, sondern eher umfangreich und komplex sind. Selbst propositionales Verstehen speist sich meinem Vorschlag zufolge aus einem holistischen Verständnis (andernfalls würden wir nicht von Verstehen, sondern z. B. von Wissen sprechen). Umfang und Komplexität typischer Verstehensobjekte machen wiederum ein hinreichend umfangreiches und komplexes Überzeugungssystem erforderlich. Ein solches Überzeugungssystem weist deutliche Parallelen zu den rechtfertigenden Relationen zwischen Überzeugungen auf, die kohärentistische Theorien postulieren: Intuitiv wird nicht nur Rechtfertigung, sondern auch das Verstehen selbst untergraben, wenn die Überzeugungen bezüglich des jeweiligen Verstehensobjekts widersprüchlich oder völlig zusammenhanglos sind. Verstehen erfordert daher ein gewisses Maß an Kohärenz zwischen den zugehörigen Überzeugungen. Jonathan Kvanvig schlägt angesichts dessen vor, dass kohärentistische Theorien womöglich besser bei Verstehen als bei Wissen aufgehoben seien (ebd., S. 193).

10.2.3 Plädoyer für einen hybriden internalistischen Ansatz

Wenn die vorherigen Überlegungen stimmig sind, erfordert Verstehen internalistische, kohärentistische Rechtfertigung. Doch ist der Kohärentismus bekanntermaßen mit einem Problem konfrontiert: dem Isolationseinwand (Feldman, 2003, S. 68–70). Kohärentist:innen nehmen üblicherweise an, dass ausschließlich Überzeugungen andere Überzeugungen rechtfertigen können (indem sie ein adäquates Netzwerk bilden). Allerdings scheinen sich Personen mit denselben Überzeugungen, aber unterschiedlichen Wahrnehmungserlebnissen in ihrer Rechtfertigung potenziell zu unterscheiden. Richard Feldmans Beispiel ist der Philosophieprofessor Magic Feldman, der sich während seiner Vorlesung vorstellt, Basketball zu spielen und exakt dieselben Überzeugungen besitzt wie der Basketballer Magic Johnson während eines Spiels (ebd., S. 68). Trotz derselben, gleichermaßen kohärenten Überzeugungen scheint Magic Johnsons Überzeugung, dass er Basketball spielt, weitaus besser gerechtfertigt als dieselbe Überzeugung von Magic Feldman. Der Vorwurf lautet also, dass der Kohärentismus keine ausreichende Verbindung zur Außenwelt (und damit

¹¹⁷ Sofern dieses Netzwerk zumindest moderat faktiv sein muss – wofür ich in Kapitel 8 argumentiert habe –, kommt natürlich ein externalistisches Element ins Spiel. Siehe dazu auch Kvanvig, 2003, S. 198 (*et passim*).

zur Wahrheit) sicherstellt. Kohärentistische Rechtfertigung verfehlt demnach ihren Zweck – nämlich dem Ausbilden wahrer Überzeugungen zuträglich zu sein.

Fundamentalistische Ansätze haben dagegen den Vorteil, dass sie üblicherweise Wahrnehmungen als fundamentale Rechtfertigung für basale Überzeugungen anerkennen und dadurch den nötigen Kontakt zwischen Welt und Überzeugung sicherstellen. Andererseits wurde in Abschnitt 10.2.2 deutlich, dass sich zentrale Annahmen kohärentistischer Ansätze in dem für Verstehen charakteristischen Begreifen von Zusammenhängen wiederfinden und sich daher besonders gut in eine Verstehenstheorie einfügen. Angesichts dieser Überlegungen ist es attraktiv, bezüglich der Rechtfertigung von Verstehensüberzeugungen weder einen starken Fundamentalismus noch einen reinen Kohärentismus zu vertreten. Stattdessen bietet sich ein hybrider Ansatz an, der kohärentistische und fundamentalistische Annahmen kombiniert. Betrachten wir also einige Vorschläge dazu, wie eine hybride Theorie der Rechtfertigung aussehen könnte.

Dualismus

Ein solcher Vorschlag, den ich der Einfachheit halber als ‚Dualismus‘ bezeichnen möchte, stammt von Eva Schmidt.¹¹⁸ Sie stellt zwei notwendige Bedingungen für die Rechtfertigung von Verstehensüberzeugungen auf: „S’s beliefs that *p*, *q*, *r*, *s* (and that *q*, *r*, *s* explain why *p*) are epistemically *justified*: they are (a) supported by *sufficient evidence* and they (b) *cohere* among each other and with S’s background beliefs“ (Schmidt, 2026, S. 230; Hervorhebung im Original).

Die erste Bedingung ist in einem evidenzialistischen Sinne fundamentalistisch: Um gerechtfertigt zu sein, müssen Verstehensüberzeugungen durch hinreichende Evidenzen gestützt werden. Diese Rechtfertigung geht letztlich zurück auf basale Überzeugungen, die keiner weiteren Rechtfertigung bedürfen. Die zweite Bedingung ist kohärentistisch: Die evidenzialistisch gerechtfertigten Überzeugungen müssen zusätzlich kohärent sein, sich also gegenseitig stützen. Rechtfertigung erfordert, dass beide Bedingungen zu einem ausreichenden Grad erfüllt sind. Allerdings werden evidenzielle und kohärentistische Rechtfertigung nicht zu einer ‚Gesamtrechtfertigung‘ gebündelt, sondern sind unabhängig voneinander. Schmidts Vorschlag fügt sich daher nicht unmittelbar in BonJours Unterscheidung zwischen starkem, moderatem und schwachem Fundamentalismus ein (siehe Abschnitt 10.2.1).

Fundhärentismus

Ein weiterer hybrider Vorschlag ist Susan Haacks (1993; 2000) ‚Fundhärentismus‘ (engl. *foundherentism*). Ihr Vorschlag heißt Fundhärentismus, weil er Rechtfertigung durch eine Kombination fundamentalistischer und kohärentistischer Elemente ausbuchstabiert. Haack erklärt das Prinzip anhand einer Analogie zu Kreuzworträtseln: Darin stützen sich die

¹¹⁸Die folgende Darstellung stützt sich neben dem zitierten Aufsatz auch auf persönliche Konversation.

einzutragenden Wörter dank ihrer Überschneidungen gegenseitig (das kohärentistische Element). Diese Stützungsrelation ist aber nicht zirkulär, weil die einzelnen Wörter zusätzlich durch die Ausfüllhinweise (z. B. ‚Erfinder des Computers‘, vier Buchstaben) gestützt werden (das fundamentalistische Element).

Die Ausfüllhinweise für eine bestimmte Lücke im Rätsel entsprechen dem Gehalt der experientiellen Evidenzen (beispielsweise introspektive oder perzeptuelle Erfahrungen) des epistemischen Subjekts. Die sich damit überschneidenden, bereits ausgefüllten Lücken entsprechen dem Inhalt der schon existierenden Überzeugungen (Haack, 2000, S. 230). Da diese Kohärenz erforderlich ist, um Überzeugungen ausreichend zu rechtfertigen, handelt es sich bei Haacks Vorschlag um eine bestimmte Form des schwachen Fundamentalismus (Van Cleve, 2014, S. 257).¹¹⁹

Überlegungsgleichgewicht

Die Grundidee des Überlegungsgleichgewichts besteht darin, dass verschiedene Annahmen mit einer gewissen Anfangsplausibilität so lange gegenseitig adaptiert, ergänzt oder revidiert werden, bis ein durch Überlegung gestütztes, kohärentes Gleichgewicht entsteht. Catherine Elgin hat dieses Prinzip unter anderem auf objektuales Verstehen übertragen.¹²⁰ Den Ausgangspunkt bilden Ausgangsverpflichtungen (engl. *antecedent commitments* oder *initially tenable commitments*). Dabei kann es sich um Intuitionen, Faustregeln oder Aberglaube, aber auch um Verpflichtungen auf bewiesene Theoreme oder andere solide Evidenzen handeln (Elgin, 2017, S. 64, vgl. auch Elgin, 2021, S. 1594). Sie bilden zusammen mit anderen Annahmen, auf die sich eine Person verpflichtet, eine Auffassung (engl. *account*) des verstandenen Themas (Elgin, 2017, S. 12–13). Damit die Verpflichtung auf eine Auffassung im Überlegungsgleichgewicht ist, müssen erstens die Elemente der Auffassung jeweils im Lichte der anderen vernünftig erscheinen. Zweitens muss die Auffassung als Ganzes – gemessen an den Ausgangsverpflichtungen – mindestens genauso vernünftig scheinen wie die verfügbaren Alternativen (ebd., S. 66). Wenn die (ausreichende) Wahrheit der Auffassung die beste Erklärung für ihre Kohärenz ist, ist die Verpflichtung auf die Auffassung gerechtfertigt (z. B. ebd., S. 71, 75).

Ein solches Überlegungsgleichgewicht ist zum einen kohärentistisch, insofern die Elemente der Auffassung konsistent sind und sich gegenseitig stützen (ebd., S. 71). Zum anderen ist es fundamentalistisch, weil manche Ausgangsverpflichtungen auch unabhängig von ihrer Einbettung in die Auffassung glaubwürdig sind, z. B. bestimmte Überzeugungen, die durch Wahrnehmung, das Zeugnis anderer oder Hintergrundtheorien gestützt werden.

¹¹⁹Haack selbst lehnt diese Kategorisierung ab. Nach gängiger Lesart beruht dies aber auf ihrer (falschen) Annahme, dass alle Fundamentalist:innen der Kohärenz jegliche Rechtfertigungskraft absprechen würden – siehe dazu Abschnitt 10.2.1.

¹²⁰Eine aktuelle Darstellung findet sich in *True Enough* (Elgin, 2017, insbesondere Kap. 4–5). Für eine Rekonstruktion und Erweiterung von Elgins Vorschlag siehe Baumberger & Brun, 2017, 2021. Den letztgenannten Aufsatz diskutiert Elgin, 2021.

Solche Ausgangsverpflichtungen erhalten diesen Status aber selbst erst durch die Einbettung in eine Auffassung (z. B. der Wahrnehmung; Elgin, 2017, S. 76–81). Elgin klassifiziert das Überlegungsgleichgewicht daher als schwach fundamentalistisch in BonJours Sinne (Elgin, 2014, S. 254).¹²¹

Probleme hybrider Theorien

Natürlich sind alle genannten hybriden Rechtfertigungstheorien mit Einwänden konfrontiert. Beispielsweise lehnt BonJour schwach fundamentalistische Theorien ab.¹²² Sein wichtigster Einwand lautet, dass unklar sei, wieso Kohärenz ein geeignetes Mittel sein solle, um manche Ausgangsverpflichtungen zu verwerfen und andere zu untermauern – schließlich gingen Fundamentalist:innen davon aus, dass Kohärenz selbst keine unabhängige Quelle von Rechtfertigung sei (BonJour, 1985, S. 29). Und wenn eine ‚Verrechnung‘ des Rechtfertigungsstatus von Ausgangsüberzeugungen und der Zunahme von Kohärenz erfolgen solle, bleibe der zugehörige ‚Umrechnungskurs‘ zu unklar, um ihn ernsthaft beurteilen zu können (ebd.). James Van Cleve spitzt BonJours Kritik weiter zu: Wenn Kohärenz dazu in der Lage sei, gewisse Überzeugungen (o. ä.) über die erforderliche ‚Rechtfertigungsschwelle‘ zu heben, dann sollte ein purer Kohärentismus allein dafür bereits ausreichen (Van Cleve, 2014, S. 257).

Weiterhin argumentiert Van Cleve, dass der von schwachen Fundamentalist:innen angenommene, geringe Rechtfertigungsgrad der Ausgangsüberzeugungen (o. ä.) nicht ausreiche, um hinreichende Rechtfertigung zu ermöglichen. Sein Vorwurf lautet, dass schwache Fundamentalist:innen zirkulär argumentierten: Ausgangsüberzeugungen sind erst durch die Kohärenz mit anderen Überzeugungen ausreichend gerechtfertigt. Doch um diese Kohärenz beurteilen zu können, braucht man gerechtfertigte Annahmen darüber, welche Überzeugungen einander stützen. Diese Annahmen leiten wir Van Cleve zufolge induktiv aus unseren Erinnerungen ab. Da solche Erinnerungen ohne Kohärenz nicht ausreichend gerechtfertigt sind, können sie das erforderliche Kohärenzmaß aber nicht begründen. Daher, so Van Cleves Ergebnis, müsse man entweder auf den Externalismus ausweichen, oder einen mindestens moderaten Fundamentalismus annehmen (ebd., S. 261–262).

Da Schmidts Dualismus die fundamentalistischen und kohärentistischen Elemente der Rechtfertigung nicht miteinander verzahnt, ist ihr Vorschlag gegen BonJours und Van Cleves Kritik gefeit. Allerdings wirft ihre Rechtfertigungsbedingung andere Fragen auf. Führen wir uns zunächst eine Auslegung ihres Vorschlags vor Augen, die Schmidt *nicht* intendiert. Dieser Lesart zufolge würde es sich um eine moderat fundamentalistische Position handeln. Basale Überzeugungen (z. B. Wahrnehmungsüberzeugungen) wären dann auch ohne Kohärenz hinreichend gerechtfertigt, doch kann sich Kohärenz positiv auf

¹²¹Eine kritische Diskussion von Elgins schwachem Fundamentalismus bietet der Austausch zwischen Elgin und James Van Cleve, zusammengestellt in Kapitel 10 von Steup et al. (2014).

¹²²Auch wenn er nicht auf diese Ansätze eingeht, lässt sich BonJours Einwand plausiblerweise auf Haacks Fundhärentismus und Elgins Überlegungsgleichgewicht übertragen.

den Rechtfertigungsstatus dieser und anderer, nicht-basaler Überzeugungen auswirken. Schmidts Vorschlag machte dann bloß eine fundamentalistische Annahme explizit, die auch in Bezug auf Wissen verbreitet ist (siehe Abschnitt 10.2.1). Entsprechend würde es sich streng genommen nicht mehr um eine hybride Theorie handeln.

Die von Schmidt intendierte Auslegung lautet stattdessen, dass kohärentistische und fundamentalistisch-evidenzielle Rechtfertigung zwei unabhängige Rechtfertigungsdimensionen darstellen. Doch dann lässt sich fragen, weshalb überhaupt zwei Bedingungen erforderlich sind. Welchen Beitrag leistet Kohärenz, wenn die einzelnen Überzeugungen dieses Netzwerks bereits individuell evidenzialistisch gerechtfertigt sind (bzw. umgekehrt)? Es scheint unstrittig, dass Verstehen ein Netzwerk an kohärenten Überzeugungen erfordert. Doch das Vorhandensein eines solchen Überzeugungssystems könnte auch als eigene Verstehensbedingung aufgefasst werden, die unabhängig von Rechtfertigung ist.

Eine überzeugende hybride Theorie der Rechtfertigung muss Antworten auf diese und viele weitere Einwände finden. Allerdings sind der reine Fundamentalismus oder Kohärentismus sowie reliabilistische bzw. tugendbasierte Ansätze gleichermaßen mit Kritik konfrontiert. Im Lichte der obigen Diskussion erscheint mir eine internalistische, hybride Theorie der Rechtfertigung am besten mit charakteristischen Eigenschaften von Verstehen übereinzustimmen. Dualismus, Fundhärentismus und Überlegungsgleichgewicht zeigen Möglichkeiten dafür auf, wie eine solche Rechtfertigungstheorie für Verstehen fruchtbar gemacht werden kann.

10.3 Instrumentelles Verstehen erfordert Rechtfertigung

Die vorherigen Abschnitte haben gezeigt, dass Verstehensüberzeugungen gerechtfertigt sein müssen. Aufgrund der spezifischen Merkmale von Verstehen ist ein hybrider, internalistischer Ansatz, der kohärentistische und fundamentalistische Elemente kombiniert, dafür besonders attraktiv. Einige Philosoph:innen bezweifeln zwar, dass Verstehen Rechtfertigung erfordert. Allerdings sind die vermeintlichen Beispiele für Verstehen ohne Rechtfertigung weit weniger plausibel, sobald man verschiedene Aspekte bzw. Objekte von Verstehen klar voneinander trennt. Doch selbst wenn man diese Argumentation nicht überzeugend findet, gibt es gute Gründe anzunehmen, dass zumindest *instrumentelles* Verstehen Rechtfertigung erfordert – nämlich einfach aufgrund des Zwecks, den Rechtfertigung nach gängiger Annahme erfüllen soll.

Die Standardauffassung nimmt eine Art Lenkungsmodell der Rechtfertigung an: Rechtfertigung soll dem epistemischen Subjekt dabei helfen, falsche Überzeugungen zu vermeiden und wahre Dinge zu glauben. Rechtfertigung ist demnach ein Mittel zur Wahrheit, das uns – in Ermangelung eines „Wahrheitssinns“ (Grundmann, 2017, S. 167) – wenigstens üblicherweise zu wahren Überzeugungen führt (BonJour, 1985, S. 7). Harrys Vorgehen in *Technik-Fan* ist der Erlangung wahrer Überzeugungen nicht zuträglich, auch wenn er durch Zufall die Wahrheit getroffen hat. Daher ist seine Überzeugung darüber, warum die

Mail seiner Mutter als ‚Spam‘ klassifiziert wurde, nicht gerechtfertigt.

Die verschiedenen Rechtfertigungstheorien konkretisieren dieses Lenkungsmodell in unterschiedlicher Form: Evidenzialist:innen nehmen an, dass durch Evidenzen gestützte Überzeugungen wahrscheinlich wahr sind; Kohärentist:innen und Anhänger:innen des Überlegungsgleichgewichts betrachten z. B. die Widerspruchsfreiheit und adäquate Vernetztheit von Überzeugungen als Wahrheitsindikator; Prozessreliabilist:innen definieren Rechtfertigung durch Berufung auf Prozesse, die der Wahrheit zuträglich sind; und vielen Tugenderkenntnistheoretiker:innen zufolge sind epistemische Tugenden solche, die dem epistemischen Subjekt zu wahren Überzeugungen verhelfen. Rechtfertigung bietet somit Leitplanken, die wahre Überzeugungen wahrscheinlich (anstatt außergewöhnlich) machen sollen. Dadurch stellt sie sicher, dass das Verstehenssubjekt seine Überzeugungen nicht so willkürlich formt wie Harry aus *Technik-Fan*, sondern vernünftig vorgeht, sich auf vertrauenswürdige und zuverlässige Quellen stützt, seine Evidenzen sorgfältig bewertet und miteinander abwägt.

Ich denke, dass der Zweck von Rechtfertigung dafür spricht, dass zumindest instrumentelles Verstehen notwendigerweise Rechtfertigung erfordert – auch unabhängig von Intuitionen bezüglich verschiedener Fallbeispiele. Instrumentelles Verstehen ist definiert als ein epistemischer Zustand, der es Stakeholdern ermöglicht, relevante gesellschaftliche Desiderata in KI-Kontexten epistemisch und ggf. auch substanziell zu erfüllen. Es ist nicht bloß im Interesse des fraglichen Stakeholders, Überzeugungen derart zu erwerben, dass sie wahrscheinlich wahr sind. Die Kopplung an ein bestimmtes Desideratum ist *konstitutiv* für instrumentelles Verstehen. Daher muss eine zuverlässige Verknüpfung zwischen Überzeugung und Desideratum in die Definition instrumentellen Verstehens integriert werden. Der naheliegendste Kandidat für die Etablierung einer solchen Verknüpfung ist Rechtfertigung: Sie bietet den Stakeholdern in XAI-Kontexten die nötigen epistemischen Leitplanken, um (wahrscheinlich) zu wahren zentralen Überzeugungen zu gelangen. Da diese Überzeugungen die Erfüllung relevanter Desiderata ermöglichen, besteht ein direkter Zusammenhang zwischen Rechtfertigung und dem Zweck instrumentellen Verstehens.

In Abschnitt 5.1.4 habe ich Rechtfertigung als eine der Dimensionen herausgestellt, anhand derer der Verstehensgrad variieren kann. Wenn instrumentelles Verstehen, wie von mir vorgeschlagen, in einem spezifischen Verstehensgrad besteht, stellt sich die Frage, welcher Grad der Rechtfertigung erforderlich für instrumentelles Verstehen ist. Ich denke, dass dies von kontextuellen Faktoren abhängt. Betrachten wir dazu einen Aspekt der jüngeren philosophischen Debatte zu Wissen. Dort wird unter dem Schlagwort des ‚pragmatischen Eingreifens‘ (engl. *pragmatic encroachment*) seit einigen Jahren die Frage diskutiert, ob kontextuelle, nicht-epistemische Faktoren (etwa praktische Ziele oder die Risiken, die mit einer falschen Überzeugung einhergehen) beeinflussen, ob ein epistemisches Subjekt in seinen Überzeugungen gerechtfertigt ist bzw. Wissen besitzt (etwa Fantl & McGrath, 2009; Stanley, 2005). Laut der These des pragmatischen Eingreifens sprechen wir einer Person z. B. Wissen zu, wenn zwar ein gewisses Risiko dafür besteht, dass ihre Überzeugung falsch ist, eine falsche Überzeugung aber keine nennenswerten negativen Konsequenzen hätte.

Ist es hingegen sehr wichtig, dass die Überzeugung wahr ist, bewerten wir die Situation des epistemischen Subjekts anders und sprechen ihr Wissen ab.

Ein Beispiel: Nehmen wir an, ich möchte heute einen Zug nach Siegen nehmen und habe dafür vor ein paar Tagen die Abfahrtszeit herausgesucht. Wenn ich in Siegen lediglich durch die Innenstadt schlendern will, ist es plausibel zu sagen, dass ich weiß, um wie viel Uhr der Zug abfährt. Es hängt nicht viel daran, ob meine Überzeugung wahr ist oder nicht – wenn ich den Zug aufgrund einer Fahrplanänderung verpasse, kann ich einfach in den nächsten Zug steigen und etwas später durch die Stadt spazieren. Habe ich hingegen ein Bewerbungsgespräch oder einen wichtigen Arzttermin, steht deutlich mehr auf dem Spiel. In einem solchen Fall scheine ich nicht zu wissen, wann der Zug abfährt, weil die mehrere Tage alte Fahrplanauskunft unzureichend ist, um meine Überzeugung zu rechtfertigen.

Auch in vielen KI-Anwendungen steht eine Menge auf dem Spiel (z. B. Medizin, autonomes Fahren, Militär oder Rechtsprechung). Vor diesem Hintergrund haben manche Autor:innen argumentiert, dass pragmatisches Eingreifen auch Verstehenszuschreibungen in XAI-Kontexten beeinflusst. Beispielsweise charakterisiert Emily Sullivan Verstehensgrade anhand der Anzahl der Was-wäre-wenn-Fragen, die ein epistemisches Subjekt beantworten kann, und geht davon aus, dass es einen kontextspezifischen Schwellwert für Verstehenszuschreibungen gibt (Sullivan, 2023, S. 315–316). Dieser Schwellwert wird wiederum davon beeinflusst, wie viel auf dem Spiel steht (ebd., Kap. 4) – entweder im Anwendungsbereich des KI-Systems (z. B. Krebserkennung versus Online-Shopping) oder individuell für die verstehende Person (z. B. eine Person, die ganz dringend einen Kredit braucht versus eine Person, die das Geld auch anders beschaffen könnte).

Wenn wir diese Überlegungen auf meine Systematisierung verschiedener Dimensionen der Gradierbarkeit von Verstehen übertragen (siehe Abschnitt 5.1), wird deutlich, dass sich Sullivans Vorschlag nicht unmittelbar auf die Rechtfertigung von Verstehensüberzeugungen auswirkt, sondern auf den Grad des Begreifens. Hohe Risiken bzw. große Chancen würden den für instrumentelles Verstehen erforderlichen Verstehensgrad anheben, weil Verstehenssubjekte in höherem Maße über die Fähigkeit verfügen müssten, Was-wäre-wenn-Fragen zu beantworten. Da ich mich in Abschnitt 9.3 dafür ausgesprochen habe, dass sich (abgesehen von grundlegenden Fähigkeiten des Erkennen-Clusters) keine für instrumentelles Verstehen notwendigen Fähigkeiten identifizieren lassen, fügt sich Sullivans Vorschlag nicht gut in meine eigenen Überlegungen ein. Mitunter scheint es für die Erfüllung eines Desideratums beispielsweise zu genügen, die Ausgabe eines KI-Systems zu erklären, auch ohne kontrafaktische Fragen beantworten zu können. Allerdings habe ich auch vorgeschlagen, dass verstehensbezogene Fähigkeiten aus einem adäquaten Überzeugungssystem erwachsen. Wenn das stimmt, bestünde ein positiver Zusammenhang zwischen der Kohärenz von Verstehensüberzeugungen, und damit auch Rechtfertigung, sowie verstehensbezogenen Fähigkeiten wie jener, die Sullivan diskutiert. Demzufolge heben hohe Risiken bzw. große Chancen den für instrumentelles Verstehen erforderlichen Grad der Kohärenz des Überzeugungssystems (d. h. der Rechtfertigung) an, was sich wiederum in einer verbesserten Fähigkeit, kontrafaktische Fragen zu beantworten, niederschlagen

kann.

Eine weitere Auseinandersetzung mit Verstehen und pragmatischem Eingreifen im KI-Kontext findet sich in dem bereits oben diskutierten Aufsatz von Eva Schmidt (2026). Sie vergleicht zwei (fiktive) Fälle: Im ersten Fall muss Luana, Mitglied eines Bewährungsausschusses für Straftäter:innen, entscheiden, ob dem Schwarzen Häftling Connor ein Kurs zur Aggressionsbewältigung angeboten wird. In ihre Entscheidung fließt auch die von COMPAS (siehe Abschnitt 3.3.2) berechnete Rückfallwahrscheinlichkeit Connors ein. Das System weist Connor eine hohe Rückfallwahrscheinlichkeit zu und erklärt die Entscheidung damit, dass viele Straftäter:innen mit ähnlichen Musikinteressen wie Connor kurz nach der Entlassung wieder straffällig werden. Diese Erklärung ist zwar überraschend für Luana, doch steht nicht viel auf dem Spiel: Falls die Erklärung falsch sein oder relevante Informationen auslassen sollte, würde Connor lediglich ein überflüssiger Kurs angeboten. Daher, so Schmidts Vorschlag, sei es intuitiv plausibel, dass Luana versteht, warum Connor eine hohe Rückfallwahrscheinlichkeit zugewiesen wurde. Im zweiten Fall steht dagegen viel auf dem Spiel: Die Richterin Juana soll über Connors Strafmaß entscheiden und berücksichtigt dafür ebenfalls die von COMPAS errechnete Rückfallwahrscheinlichkeit. Auch hier wird Connor ein hohes Rückfallrisiko zugewiesen, und Juana erhält dieselbe, gleichermaßen überraschende Erklärung für diese Entscheidung wie Luana. Schmidt zufolge sind wir im zweiten Fall nicht geneigt, Juana Verstehen zuzugestehen, denn hier drohen signifikante negative Konsequenzen. Es könnte beispielsweise sein, dass COMPAS den Musikgeschmack als Proxy für Connors ethnische Zugehörigkeit verwendet und Juana ein diskriminierendes oder unverhältnismäßig harsches Urteil fällt, wenn sie der Erklärung folgt (ebd., S. 227).

Der Kontext greift in diesen Fällen pragmatisch in beide Bedingungen von Rechtfertigung ein, die Schmidt aufstellt (siehe Abschnitt 10.2.3). Zur Erinnerung: Damit das Überzeugungsnetzwerk eines Subjekts gerechtfertigt ist, müssen Schmidt zufolge die zugehörigen Überzeugungen hinreichend durch Evidenzen gestützt werden. Zusätzlich muss das Überzeugungssystem in sich kohärent sein und mit relevanten Hintergrundüberzeugungen kohärieren (ebd., S. 230). Wenn viel auf dem Spiel steht, sorgt dies zum einen dafür, dass für Rechtfertigung mehr Evidenzen benötigt werden; zum anderen führt dies dazu, dass für Rechtfertigung ein höherer Grad an Kohärenz zwischen den einzelnen Überzeugungen bestehen muss. Und da Verstehen laut Schmidts Vorschlag Rechtfertigung erfordert, können pragmatische Faktoren beeinflussen, ob eine Person in einem bestimmten Kontext Verstehen besitzt oder nicht.

Übertragen auf instrumentelles Verstehen heißt das Folgendes: Sobald für den fraglichen Stakeholder viel auf dem Spiel steht, müssen seine Überzeugungen über relevante systembezogene Aspekte besser gerechtfertigt sein, um die für instrumentelles Verstehen erforderliche Rechtfertigungsschwelle zu überschreiten. Wenn wenig auf dem Spiel steht, genügt bereits ein geringerer Grad der Rechtfertigung.¹²³

¹²³ Allerdings kritisiert Schmidt dieses Vorgehen in einem anderen Aufsatz (Schmidt, 2025). Anstatt anzunehmen,

Es ist außerdem plausibel, dass zentrale Überzeugungen besser gerechtfertigt sein müssen als periphere. In Abschnitt 7.4 habe ich vorgeschlagen, dass vornehmlich das jeweils relevante Desideratum des instrumentell verstehenden Stakeholders festlegt, welche instrumentellen Verstehensüberzeugungen zentral und welche peripher sind. Zentrale Überzeugungen sind wichtiger für die Erfüllung des jeweiligen Desideratums und damit für das Erreichen des Zwecks von instrumentellem Verstehen. Da zentrale Überzeugungen wahr oder zumindest näherungsweise wahr sein müssen, und solide Rechtfertigung die Wahrheit einer Überzeugung wahrscheinlicher macht, müssen zentrale Überzeugungen besonders gut gerechtfertigt sein – insbesondere dann, wenn viel auf dem Spiel steht. Das spricht dafür, dass pragmatisches Eingreifen kontextuell zentrale Überzeugungen in höherem Maße betrifft.

Nichtsdestotrotz müssen auch periphere Überzeugungen gerechtfertigt sein. Ein gewisser Grad der kohärentistischen Rechtfertigung ergibt sich bereits daraus, dass auch periphere Überzeugungen Teil des instrumentellen Überzeugungssystems sind und daher mit anderen Überzeugungen über systembezogene Aspekte kohärieren. Um dem Isolationsseinwand zu entgehen, braucht es darüber hinaus aber auch die fundamentalistische Rechtfertigungsdimension.

Damit schließe ich meine Diskussion der letzten von mir ausgewählten Subdebatte zu Verstehen ab. Im nun folgenden Kapitel bündele ich die bisher gewonnenen Einsichten und führe meinen Begriff instrumentellen Verstehens ein.

dass hohe Risiken bzw. große Chancen den Schwellwert anheben, der für hinreichende Rechtfertigung erforderlich ist, sei es plausibler, dass diese die vorhandenen normativen Gründe so weit abschwächen, dass sie die Überzeugung nicht länger rechtfertigen. Der Schwellwert für Rechtfertigung bleibt dadurch unverändert – hohe Risiken bzw. große Chancen sorgen lediglich dafür, dass dieser Schwellwert nicht überschritten wird, weil die Gründe zu schwach sind.

11 Instrumentelles Verstehen

Aufbauend auf der Argumentation der vorherigen Kapitel formuliert dieses Kapitel eine Antwort auf die erste Forschungsfrage, die ich in Abschnitt 1.1 aufgeworfen habe:

- (a) Welche Form des Verstehens ermöglicht Stakeholdern die Erfüllung gesellschaftlicher Desiderata in XAI-Kontexten?

Die Antwort lautet natürlich: instrumentelles Verstehen. Um diesen epistemischen Zustand näher zu bestimmen, habe ich mich für einen begriffsgestaltenden Ansatz entschieden, der sich an Rudolf Carnaps Methode der Explikation orientiert (Kapitel 4). Das von Georg Brun formulierte ‚Rezept fürs Explizieren‘ (Abschnitt 4.1.3) zerlegt den Prozess der Begriffsgestaltung in vier Schritte: Der erste Schritt erfolgte in Abschnitt 4.2. Er bestand in der Spezifikation der Explikationsaufgabe, einschließlich der Präzisierung des Explikandums (Schritt 1a), der Benennung des theoretischen Rahmens (1b) und der Zwecke, die das Explikat in diesem Rahmen erfüllen soll (1c). Der zweite Schritt, nämlich die Formulierung von Adäquatheitsbedingungen für das Explikat, folgte direkt im Anschluss in Abschnitt 4.3. Dort habe ich sowohl Ähnlichkeits- (Schritt 2a) als auch Nützlichkeitsbedingungen (2b) formuliert, die das Explikat erfüllen muss, damit die Explikation als erfolgreich gelten kann. In den vorangegangenen Kapiteln 5 bis 10 habe ich mich mit erkenntnistheoretischen Debatten zu Verstehen auseinandergesetzt, um sicherzustellen, dass das Explikat diesen Adäquatheitsbedingungen gerecht wird.

In diesem Kapitel schließe ich die Explikation ab. Im nun folgenden Abschnitt 11.1 erfolgt Schritt (3) der Begriffsgestaltung, nämlich die Einführung des Explikats in Form meiner Definition instrumentellen Verstehens. In Abschnitt 11.2 findet abschließend Schritt (4) statt: Ich evaluiere mein Explikat anhand der zuvor formulierten Adäquatheitsbedingungen.

11.1 Einführung des Explikats

Der dritte Schritt des Rezepts besteht in der Einführung des Explikats. Was ist also instrumentelles Verstehen? Rufen wir uns noch einmal in Erinnerung, was in den vorangegangenen Kapiteln erarbeitet worden ist:

- *Instrumentelles Verstehen ist nicht gradierbar (Abschnitt 5.2).* Aufgrund seiner Zweckgebundenheit ist instrumentelles Verstehen von binärer Qualität. Allerdings entspricht instrumentelles Verstehen einem bestimmten Verstehensgrad entlang der

Faktizitäts-, Begreifens- und Rechtfertigungsdimension (vgl. Abbildung 5.1). Der Schwellwert, den ein Stakeholder hinsichtlich dieser Dimensionen überschreiten muss, um instrumentell zu verstehen, wird durch kontextuelle Faktoren bestimmt.

- *Die Objekte instrumentellen Verstehens sind komplex (Abschnitt 6.3).* Die Objekte instrumentellen Verstehens sind verschiedene systembezogene Aspekte und weisen ein Mindestmaß an Komplexität auf. Instrumentelles Verstehen kann unterschiedliche Repräsentationsebenen betreffen, also symbolisch oder faktisch sein. Der hier vorgeschlagenen Lesart zufolge handelt es sich typischerweise um faktisches Verstehen systembezogener Aspekte, vermittelt durch das symbolische Verstehen eines Verstehensermöglichers.
- *Instrumentelles Verstehen erfordert Überzeugung (Abschnitt 7.4).* Die in instrumentelles Verstehen involvierte epistemische Einstellung ist Überzeugung. Diese Bedingung ist mit symbolischem und symbolisch vermitteltem faktischen Verstehen vereinbar und auch damit, dass Verstehensobjekte (insbesondere Verstehensermöglicher) nicht zwangsläufig propositional sind.
- *Instrumentelles Verstehen ist moderat faktiv (Abschnitt 8.4).* Instrumentelles Verstehen ist mit falschen Überzeugungen kompatibel, solange die zentralen Überzeugungen wahr oder näherungsweise wahr sind. Die Zentralität einer Verstehensüberzeugung wird vornehmlich durch das Desideratum des instrumentell verstehenden Stakeholders bestimmt. Die Rolle dieses Stakeholders kann dabei helfen, zentrale Überzeugungen auf ein realistisches Maß einzugrenzen.
- *Instrumentelles Verstehen erfordert Fähigkeiten (Abschnitt 9.3).* Eine instrumentell verstehende Person verfügt über verstehensbezogene Fähigkeiten, die sich den sechs Clustern erkennen, bewerten, erklären, schlussfolgern, intervenieren und/oder entwickeln zuordnen lassen. Welche Fähigkeiten jeweils erforderlich für instrumentelles Verstehen sind, hängt vom Desideratum des fraglichen Stakeholders ab, ggf. unter Berücksichtigung seiner Rolle.
- *Instrumentelles Verstehen erfordert Rechtfertigung (Abschnitt 10.3).* Die in instrumentelles Verstehen involvierten Überzeugungen müssen gerechtfertigt sein. Eine internalistische, hybride Theorie der Rechtfertigung, die fundamentalistische und kohärentistische Elemente kombiniert, ist dafür besonders attraktiv. Der für instrumentelles Verstehen erforderliche Grad der Rechtfertigung wird durch die kontextuellen Risiken bzw. Chancen bestimmt.

Nachfolgend präzisiere ich die in den vorherigen Kapiteln erarbeiteten Merkmale instrumentellen Verstehens weiter: In Abschnitt 11.1.1 stelle ich meine Definition instrumentellen Verstehens vor. Anschließend wende ich meinen Definitionsvorschlag in Abschnitt 11.1.2

auf mehrere Beispiele an, um die Kontextsensitivität der einzelnen Bedingungen zu illustrieren. Schließlich gehe ich in Abschnitt 11.1.3 darauf ein, ob Wissen notwendig bzw. hinreichend für (instrumentelles) Verstehen ist.

11.1.1 Die Definition

Auf Grundlage der zuvor herausgearbeiteten Merkmale meines Explikats schlage ich vor, dass *instrumentelles Verstehen* durch eine Überzeugungs- und Komplexitätsbedingung, eine Objekt-, Faktizitäts-, Rechtfertigungs- und Fähigkeitsbedingung sowie durch den jeweiligen Einfluss kontextueller Faktoren bestimmt werden kann. Diese Anforderungen lassen sich in Form zweier kontextsensitiver, notwendiger und zusammen hinreichender Bedingungen ausdrücken. Die erste Bedingung verlangt vom instrumentell verstehenden Stakeholder ein gerechtfertigtes, moderat faktives Überzeugungssystem über systembezogene Aspekte. Die zweite Bedingung erfordert das Verfügen über verstehensbezogene Fähigkeiten. Hier ist die Definition:

Instrumentelles Verstehen: Für Kontext K – bestehend aus (i) einem Stakeholder S , (ii) der Rolle R von S , (iii) dem Desideratum D , das S in Bezug auf ein KI-System und/oder das übergeordnete soziotechnische System besitzt sowie (iv) den mit der (Nicht-)Erfüllung von D verbundenen Chancen und Risiken CR – gilt:
 S versteht instrumentell¹²⁴ genau dann, wenn die folgenden Bedingungen erfüllt sind:

- (1) S besitzt ein Überzeugungssystem, das die folgenden Bedingungen erfüllt:
(Überzeugungsbedingung und Komplexitätsbedingung)
 - (a) Die Inhalte der darin enthaltenen Überzeugungen sind Propositionen über systembezogene Aspekte. *(Objektbedingung)*
 - (b) Die darin enthaltenen zentralen, d. h. für die Erfüllung von D relevanten und ggf. gemäß R eingegrenzten, Überzeugungen sind zumindest annähernd wahr. *(Faktizitätsbedingung)*
 - (c) Die darin enthaltenen Überzeugungen sind *ultima facie* gerechtfertigt, wobei der dafür erforderliche Grad der Rechtfertigung von CR abhängt. *(Rechtfertigungsbedingung)*
- (2) S verfügt über für die Erfüllung von D relevante und ggf. gemäß R eingegrenzte verstehensbezogene Fähigkeiten. *(Fähigkeitsbedingung)*

¹²⁴Verstehensdefinitionen folgen in der Regel dem Schema ‚ S versteht X genau dann, wenn...‘, benennen also das Verstehensobjekt im Vordersatz der Definition. Im Gegensatz dazu wird das Objekt instrumentellen Verstehens erst in Bedingung (1a) bestimmt. Das hat den folgenden Grund: Für instrumentelles Verstehen ist entscheidend, was man mit diesem Verständnis anfangen kann (nämlich die Erfüllung des fraglichen Desideratums). Was dafür instrumentell verstanden werden muss, hängt von kontextuellen Faktoren ab. Das Objekt instrumentellen Verstehens ist somit Teil der pragmatischen, durch das Definieren zu spezifizierenden Aspekte.

Ich gehe nun kurz genauer auf diese beiden Bedingungen und die drei Subbedingungen ein, beschränke mich aber jeweils auf allgemeine Anmerkungen. Wie sich die einzelnen Bedingungen und die dafür relevanten Kontextfaktoren in konkreten Beispielen niederschlagen, verdeutliche ich in Abschnitt 11.1.2.

Bedingung (1): die Überzeugungs- und Komplexitätsbedingung

Die erste Bedingung spezifiziert das Überzeugungssystem einer instrumentell verstehenden Person. Sie trägt mehreren Einsichten aus den vorherigen Kapiteln Rechnung. Zum einen verlangt Bedingung (1) ein *Überzeugungssystem*, da sich Überzeugung als attraktivster Kandidat für die in instrumentelles Verstehen involvierte epistemische Einstellung erwiesen hat (Abschnitt 7.4). Zum anderen fordert Bedingung (1) ein Überzeugungssystem. Dadurch stellt sie in Kombination mit Subbedingung (1a) sicher, dass instrumentelle Verstehensobjekte umfangreich bzw. komplex genug sind, um das für Verstehen notwendige Begreifen von Zusammenhängen zu ermöglichen (Abschnitt 6.3). Simples ‚instrumentelles Wissen‘ einzelner Propositionen wird entsprechend ausgeschlossen. Um nicht als lose Ansammlung von Überzeugungen zu gelten, müssen die im Überzeugungssystem enthaltenen Überzeugungen außerdem kohärent sein (Abschnitt 10.2.1). Kohärenz erfordert neben Widerspruchsfreiheit auch die adäquate Verknüpftheit der Überzeugungen untereinander. Zusammengenommen stellt Bedingung (1) sicher, dass eine instrumentell verstehende Person nicht bloß einzelne, voneinander isolierte Überzeugungen besitzt, sondern ein hinreichend komplexes, vernetztes Überzeugungssystem.

Ich bin dafür offen, dass unterschiedliche Überzeugungssysteme für instrumentelles Verstehen in Frage kommen. Beispielsweise ist vorstellbar, dass eine Person die Fairness eines KI-Systems abschätzen kann, indem sie die Trainingsdaten und den Lernalgorithmus inspiziert, oder indem sie XAI-Erklärungen heranzieht, oder indem sie die Ausgaben des Systems statistisch auswertet. Plausiblerweise beeinflusst auch die Rolle eines Stakeholders, welche systembezogenen Aspekte er betrachtet bzw. welche Verstehensermöglicher er nutzt. Je nachdem, welchen Weg ein Stakeholder wählt, unterscheidet sich sein Überzeugungssystem. Diese unterschiedlichen Überzeugungssysteme machen dann jeweils ein imperfektes Verständnis der Fairness des KI-Systems aus. Instrumentelles Verstehen muss wiederum nicht maximalem Verstehen entsprechen. Sofern die einzelnen Instanzen imperfekten Verstehens jeweils die Erfüllung des relevanten Desideratums ermöglichen, spricht nichts dagegen, dass instrumentelles Verstehen durch verschiedene Überzeugungssysteme realisiert werden kann.

Bedingung (1a): die Objektbedingung Bedingung (1a) bestimmt den Inhalt der instrumentellen Verstehensüberzeugungen. Wie in Abschnitt 6.3 geschildert, sind die Objekte instrumentellen Verstehens systembezogene Aspekte, die ein KI-System direkt oder auch nur indirekt betreffen. Beispielsweise kann es sich dabei um die Ausgaben des fraglichen KI-Systems und die dafür ausschlaggebenden Aspekte der Eingabedaten handeln, aber

auch um Hintergrundinformationen, etwa zur Anwendungsdomäne und Maschinellern, oder um Aspekte des übergeordneten soziotechnischen Systems, z. B. rechtliche Rahmenbedingungen oder institutionelle Strukturen. Potenziell relevante systembezogene Aspekte umfassen also sowohl das fragliche KI-System (bzw. einzelne seiner Aspekte) als auch das übergeordnete soziotechnische System (bzw. einzelne seiner Aspekte), in welches das KI-System eingebettet ist. Da Überzeugungen auf Propositionen gerichtet sind, besteht der Inhalt der in der ersten Bedingung spezifizierten Überzeugungen in Propositionen über diese systembezogenen Aspekte.

Die Überzeugungen über diese Aspekte können sehr umfangreich sein. Sofern sie das relevante KI-System direkt oder indirekt betreffen und miteinander verknüpft sind, zählen sie zum instrumentellen Überzeugungssystem des Stakeholders. Entsprechend tragen sie auch zur Kohärenz des Überzeugungssystems und damit zur in (1c) beschriebenen Rechtfertigung bei. Allerdings müssen nicht alle der in (1a) bestimmten Überzeugungen wahr sein – das bringt uns zur Faktizitätsbedingung.

Bedingung (1b): die Faktizitätsbedingung Bedingung (1b) bestimmt den erforderlichen Grad der Faktizität. Instrumentelles Verstehen ist moderat faktiv (siehe Abschnitt 8.4), das heißt, dass lediglich die zentralen Überzeugungen des Überzeugungssystems wahr oder zumindest annähernd wahr sein müssen. Instrumentelles Verstehen ist also damit kompatibel, dass eine Teilmenge der in (1a) bestimmten Überzeugungen falsch ist – solange diese peripher sind. Dafür muss man periphere von zentralen Überzeugungen unterscheiden. Mein Vorschlag lautet, dass die Zentralität einer instrumentellen Verstehensüberzeugung vom Kontext abhängt (Abschnitt 8.3.3) – genauer, vom Desideratum des instrumentell verstehenden Stakeholders und ggf. auch von seiner Rolle im jeweiligen KI-Kontext (etwa als Systementwicklerin, Entscheidungssubjekt oder Mensch in der Schleife).

Das Desideratum wirkt sich auf die Zentralität aus, weil manche falschen Überzeugungen die Erfüllung eines Desideratums stärker gefährden als andere. Zentrale Überzeugungen betreffen systembezogene Aspekte, die relevant für die Erfüllung des fraglichen Desideratums sind. ‚Relevant‘ heißt hier, dass diese Aspekte Aufschluss über den Status der Desideratumserfüllung geben können. Um beispielsweise zu verstehen, ob ein KI-System Datenschutzanforderungen erfüllt, braucht man wahre Überzeugungen über andere Aspekte als wenn es darum geht, die Effizienz des Systems zu beurteilen.

Die jeweils relevanten systembezogenen Aspekte werden jedoch schnell zu umfangreich, um für ein Individuum noch bewältigbar zu sein. In Abschnitt 4.2.1 habe ich daher die Rolle des fraglichen Stakeholders als limitierenden Faktor vorgeschlagen. Denn aus dieser Rolle ergeben sich berechnete Erwartungen hinsichtlich seiner Expertise: Beispielsweise ist es legitim, von einer Programmiererin mehr und detailliertere wahre Überzeugungen darüber zu erwarten, wie ein gegebenes KI-System funktioniert, als von einer Endnutzerin, die dasselbe System zur Bewältigung beruflicher Aufgaben verwendet. Letztere sollte sich hingegen – *qua* ihrer Rolle – besser mit der Anwendungsdomäne auskennen als

die Programmiererin. Meinem Vorschlag zufolge fungieren solche Expertiseerwartungen als Heuristik: Sie helfen dabei, die zentralen Überzeugungen auf ein realistisches Maß einzugrenzen. Zumindest annähernd wahre Überzeugungen hinsichtlich dieser reduzierten Aspekte sind dann notwendig für instrumentelles Verstehen.

Es kann sein, dass die zentralen Überzeugungen einer bestimmten Instanz instrumentellen Verstehens für sich genommen bereits umfangreich und kohärent genug sind, um der Komplexitätsbedingung gerecht zu werden. In solchen Fällen können die zentralen Überzeugungen prinzipiell das gesamte Überzeugungssystem eines instrumentell verstehenden Stakeholders ausmachen. Die Überzeugungen aus (1a) und (1b) fielen dann zusammen und die betreffende Instanz instrumentellen Verstehens wäre vollkommen faktiv. Selbstverständlich erfordert instrumentelles Verstehen nicht, dass ein Stakeholder falsche Überzeugungen über systembezogene Aspekte besitzt. Entscheidend ist, dass bestimmte falsche (nämlich periphere) Überzeugungen instrumentelles Verstehen nicht verhindern. Dadurch soll vermieden werden, dass instrumentelles Verstehen eine derart anspruchsvolle epistemische Errungenschaft darstellt, dass sie zumindest von Nichtfachleuten kaum oder gar nicht erlangt werden kann.

Bedingung (1c): die Rechtfertigungsbedingung Die letzte Subbedingung von Bedingung (1) verlangt, dass die Überzeugungen des Überzeugungssystems *ultima facie* gerechtfertigt sind, der fragliche Stakeholder darf also keine unangefochtenen Anfechtungsgründe für seine Überzeugungen besitzen (siehe Abschnitt 10.3). Die relevante Form der Rechtfertigung ist internalistisch (wenngleich Bedingung (1b) ein externalistisches Element enthält) und hybrid: Das Überzeugungssystem muss hinreichend kohärent sein und zudem ein Fundament an basalen Überzeugungen aufweisen, das die nicht-basalen Überzeugungen stützt (siehe Abschnitt 10.2.3). Kohärenz ist wiederum ein wichtiger Bestandteil der Begreifenskomponente instrumentellen Verstehens.

Wie gut die Überzeugungen des Überzeugungssystems gerechtfertigt sein müssen, um den für instrumentelles Verstehen erforderlichen Verstehensgrad zu erreichen, hängt abermals vom Kontext ab. Der wichtigste Kontextfaktor ist hierbei, wie viel auf dem Spiel steht – konkret, wie viel daran hängt, dass das Desideratum des Stakeholders erfüllt wird. Wenn es um die faire Vergabe eines Bankkredits mithilfe von KI geht, steht deutlich mehr auf dem Spiel als wenn ein System für die akkurate Empfehlung neuer Produkte in einem Onlineshop eingesetzt wird. In Abschnitt 10.3 habe ich vorgeschlagen, dass instrumentelles Verstehen einen höheren Rechtfertigungsgrad erfordert, wenn die Risiken hoch bzw. die Chancen groß sind – das gilt insbesondere für die zentralen Überzeugungen des Überzeugungssystems.

Bedingung (2): die Fähigkeitsbedingung

Die zweite und letzte Bedingung spezifiziert die verstehensbezogenen Fähigkeiten, die ein instrumentell verstehendes Subjekt besitzt. Diese Fähigkeiten bilden zusammen mit

der Komplexität und Kohärenz von Verstehensüberzeugungen die Begreifenskomponente instrumentellen Verstehens. Auch wenn ich es plausibel finde, dass diese Fähigkeiten aus einem adäquat verknüpften Überzeugungssystem erwachsen, sind sie doch von diesem Überzeugungssystem verschieden und daher in eine eigene Bedingung ausgelagert.

Auch die Fähigkeitsbedingung ist kontextsensitiv. Wie auch bei Bedingung (1b) bildet das Desideratum des Stakeholders den entscheidenden Kontextfaktor, und auch hier kann seine Rolle dabei helfen, diese Fähigkeiten auf ein realistisches Maß einzugrenzen. Das Desideratum wirkt sich deswegen auf die Fähigkeitsbedingung aus, weil die Erfüllung gesellschaftlicher Desiderata jeweils durch unterschiedliche verstehensbezogene Fähigkeiten ermöglicht wird. Die Rolle ist erneut deswegen relevant, weil die erforderlichen Fähigkeiten als Teil der Expertise eines Stakeholders mithilfe legitimer Expertiseerwartungen eingegrenzt werden können.

Zum Kontext

Ich habe oben bereits den Einfluss verschiedener Kontextfaktoren auf die Bedingungen instrumentellen Verstehens geschildert. Ich möchte nun einige übergeordnete Anmerkungen zum Kontext ergänzen. Erstens: Ich habe drei Kontextfaktoren diskutiert, nämlich das Desideratum des Stakeholders, seine Rolle sowie die kontextuellen Risiken bzw. Chancen. Es ist denkbar, dass weitere relevante Kontextfaktoren existieren. Ein möglicher Faktor dieser Art ist großer Zeitdruck (Langer et al., 2021, S. 10): Der Erwerb von Verstehen kostet Zeit. Wenn ein Stakeholder sehr schnell handeln muss (z. B. wenn eine Ärztin innerhalb weniger Minuten über die geeignete Behandlung einer Patientin zu entscheiden hat), genügt womöglich ein geringerer Verstehensgrad, um die kontextsensitive Schwelle instrumentellen Verstehens zu überschreiten.

Zweitens: Meine Definition bezieht sich auf ein einzelnes gesellschaftliches Desideratum. Es ist davon auszugehen, dass viele Stakeholder mehrere Desiderata besitzen. Allerdings lassen sich die jeweils erforderlichen Überzeugungssysteme und Fähigkeiten getrennt voneinander charakterisieren und dann zu einem ‚Gesamtverstehen‘ zusammensetzen. Das scheint unproblematisch zu sein, da – soweit ich sehen kann – keine Zielkonflikte zwischen der epistemischen Erfüllung verschiedener Desiderata existieren. Zwar kann die substanzielle Erfüllung einzelner Desiderata in Konflikt zueinander stehen (womöglich zählen dazu Leistungsfähigkeit und Transparenz des Systems, siehe Abschnitt 3.2.1). Doch steht der epistemische Zugang dazu, ob ein Desideratum erfüllt ist oder nicht, dem epistemischen Zugang zum Status eines anderen Desideratums nicht im Wege. Dasselbe scheint bei Desiderata zu gelten, deren epistemische und substanzielle Erfüllung zusammenfallen (siehe Abschnitt 4.2.3). Zur Erinnerung: Das betrifft Desiderata, die selbst epistemischer Natur sind – etwa menschliche Beaufsichtigung, informierte Einwilligung in KI-gestützte Entscheidungen oder auch die epistemische Bedingung moralischer Verantwortlichkeit. Es scheint unproblematisch anzunehmen, dass man systembezogene Aspekte zugleich ausreichend verstehen kann, um KI-gestützte Entscheidungen zu beaufsichtigen, informiert

in diese Entscheidungen einzuwilligen und die Verantwortung dafür zu tragen.

Drittens kann instrumentelles Verstehen mitunter eine ganz bestimmte wahre Überzeugung erfordern. Sie betrifft den Status der Erfüllung des fraglichen Desideratums, daher nenne ich sie ‚Status-Überzeugung‘. Ihr Inhalt besteht in der Proposition, dass dieses Desideratum (zu einem kontextuell ausreichenden Grad) erfüllt bzw. nicht erfüllt ist. Eine solche Status-Überzeugung muss durch das Überzeugungssystem aus Bedingung (1) inferenziell gerechtfertigt sein. Die Idee ist hier, dass der instrumentell verstehende Stakeholder aus seinem Verständnis systembezogener Aspekte explizit eine (wahre) Überzeugung darüber ableitet, ob das relevante Desideratum substanziell erfüllt ist oder nicht. Eine solche Überzeugung scheint in manchen XAI-Kontexten erforderlich zu sein. Beispielsweise könnte die Mitarbeiterin einer Zertifizierungsstelle für KI-Systeme ausdrücklich beurteilen wollen, ob ein gegebenes System bestimmte Anforderungen hinsichtlich Genauigkeit, Fairness, Vertrauenswürdigkeit, Sicherheit usw. erfüllt. Bei anderen Stakeholdern bleiben solche gesellschaftlichen Desiderata hingegen eher implizit. Angesichts dessen wäre es zu anspruchsvoll, von allen Stakeholdern eine wahre Status-Überzeugung zu verlangen.¹²⁵ Das Beispiel der Mitarbeiterin der Zertifizierungsstelle legt nahe, dass sich abermals die Rolle des Stakeholders darauf auswirkt, ob die Status-Überzeugung zu den für instrumentelles Verstehen zentralen Überzeugungen zählt.

11.1.2 Anwendung auf Beispiele

In diesem Abschnitt möchte ich exemplarisch aufzeigen, wie sich die Bedingungen instrumentellen Verstehens in konkreten Kontexten niederschlagen. Dafür stütze ich mich auf Beispielfälle aus früheren Kapiteln. Beginnen wir mit *Bewerbung* aus Abschnitt 8.4:

Bewerbung: Programmiererin Lina ist die Chefentwicklerin eines Empfehlungssystems namens *JobAI*, das bei der Personalauswahl eingesetzt wird. In das KI-System werden Informationen aus den Bewerbungsunterlagen der jeweiligen Bewerber:innen eingespeist – zum Beispiel der Lebenslauf, Zertifikate über zusätzliche Qualifikationsmaßnahmen oder Arbeitszeugnisse. Anhand dieser Informationen ordnet *JobAI* allen Bewerber:innen jeweils einen Wert entsprechend der vom System abgeschätzten Eignung zu. Der Personalmanager Fritz, der in einem anderen Unternehmen arbeitet, verwendet dieses System, um geeignete Kandidat:innen für eine zu besetzende Stelle auszuwählen. *JobAI* vergibt an Bewerberin Magda nur 3 von 10 möglichen Punkten, während Fritz Magda als besonders vielversprechende Kandidatin einschätzt.

Um möglichst unterschiedliche Instanzen instrumentellen Verstehens abzudecken, können wir den Kontext der Stakeholder Lina und Fritz so ausgestalten, dass beide jeweils ein anderes Desideratum besitzen und eine andere Rolle einnehmen.

¹²⁵Danke an Eva Schmidt, dass sie mich vom hohen Anspruch der Status-Überzeugung überzeugt hat.

Linas instrumentelles Verstehen

Beginnen wir mit Programmiererin Lina, die als Chefentwicklerin von *JobAI* maßgeblich für dessen Programmierung, Entwicklung und Wartung zuständig ist. Nehmen wir an, dass Lina erfahren hat, dass das System nicht fehlerfrei funktioniert: In einer signifikanten Anzahl von Fällen weist es geeigneten Bewerber:innen eine niedrige Punktzahl zu und schätzt weniger qualifizierte Personen zu gut ein. Linas Desideratum besteht darin, diesen Fehler zu beseitigen, und es wäre substanziell erfüllt, wenn das System frei von Fehlern ist. Doch um es epistemisch zu erfüllen, muss Lina verstehen, woher der Fehler rührt und wie sie ihn beheben kann. Es ist diese epistemische Desideratumserfüllung, die durch instrumentelles Verstehen adressiert wird.

Wir können stipulieren, dass *JobAI* während des Trainings gelernt hat, einen Proxy des Geschlechts der Bewerber:innen für deren Beurteilung heranzuziehen (vgl. Abschnitt 8.4). Wir können weiterhin annehmen, dass die falsche Einstufung mancher Bewerber:innen von dieser Verzerrung herrührt – sagen wir von bestimmten Hobbys, die in manchen Lebensläufen angegeben und häufiger von Männern bzw. Frauen ausgeübt werden. Betrachten wir nun die Bedingungen von Linas instrumentellem Verstehen im Einzelnen.

Linas Bedingung (1a) Die erste Bedingung spezifiziert Linas Überzeugungssystem hinsichtlich seiner Inhalte, Faktizität und Rechtfertigung. Aus inhaltlicher Perspektive braucht Lina laut Bedingung (1a) ein Überzeugungssystem über systembezogene Aspekte, also über Aspekte, die *JobAI* direkt und/oder indirekt betreffen. Diese Aspekte sind potenziell sehr umfangreich. Bei Systementwicklerin Lina könnte das Überzeugungssystem miteinander zusammenhängende Überzeugungen über Entwicklungshistorie und informatische Grundlagen des Systems, seine Stärken und Schwächen, seinen ungefähren Marktwert, den Beitrag verschiedener Kolleg:innen zum Endprodukt, Konkurrenzprodukte anderer Hersteller, die für *JobAI* verwendeten Trainingsdaten und vieles mehr enthalten. Insofern diese Aspekte *JobAI* direkt oder indirekt betreffen und miteinander verknüpft (anstatt isoliert voneinander) sind, zählen sie zu Linas instrumentellem Überzeugungssystem. Das ist deswegen wichtig, weil sich Überzeugungen, die Linas Überzeugungssystem angehören, auf Faktizität und Rechtfertigung ihrer Verstehensüberzeugungen auswirken.

Linas Bedingung (1b) Laut Bedingung (1b) muss Linas Überzeugungssystem moderat faktiv sein, das heißt, dass die darin enthaltenen zentralen Überzeugungen zumindest annähernd wahr sein müssen. Periphere Überzeugungen, die ebenfalls Teil des Überzeugungssystems sind, dürfen dagegen falsch sein. Der wichtigste Kontextfaktor für die Bestimmung der Zentralität ist Linas Desideratum (also die Beseitigung des Fehlers). Aufgrund ihres Desideratums braucht Lina wahre Überzeugungen über die Ursache des Fehlers, um instrumentell zu verstehen. Sie muss also wahrheitsgemäß glauben, dass die falsche Einstufung mancher Bewerber:innen durch die (übermäßige) Berücksichtigung ihrer Hobbys verursacht wird, und dass es sich dabei um einen Proxy für deren Geschlecht handelt.

Außerdem muss sie Überzeugungen über geeignete Schritte besitzen, um diesen Fehler zu beheben. Wir können stipulieren, dass sich die Verzerrung von *JobAI* beseitigen lässt, indem das System mit zusätzlichen oder besser aufbereiteten Daten trainiert wird. Überzeugungen wie diese sind für die (epistemische) Erfüllung von Linas Desideratum relevant und daher zentral: Sind sie nicht einmal annähernd wahr, bleibt Linas Desideratum unerfüllt – Lina versteht nicht, warum *JobAI* mitunter geeignete Bewerber:innen als unqualifiziert einstuft und was sie tun müsste, um dieses Problem zu beseitigen.

Zudem ist es aufgrund ihrer Rolle angemessen, von Lina eine solide technische Expertise hinsichtlich *JobAI* zu erwarten. Natürlich irren sich auch Expert:innen hin und wieder, sie braucht also keine perfekte Expertise. Doch als Chefentwicklerin sollte sich Lina sehr gut mit dem Empfehlungssystem auskennen. Die oben geschilderten wahren Überzeugungen über die Ursache des Fehlers und Strategien zu seiner Beseitigung können wir daher legitimerweise von Lina erwarten. Für Personalmanager Fritz wären solche Überzeugungen hingegen zu viel verlangt.

Zur Klarstellung: Es kann sein, dass Lina diese Überzeugungen fehlen, obwohl sie die gebotene Expertise besitzt – vielleicht ist *JobAI* derart opak, dass sogar Expert:innen davor kapitulieren müssen. Der Punkt ist, dass Lina in diesem Fall kein instrumentelles Verstehen besitzt. Die rollenbasierte Expertiseerwartung verlangt nicht zwangsläufig, dass Lina instrumentell versteht, sondern hilft dabei, die für die Erfüllung des jeweiligen Desideratums relevanten Überzeugungen derart einzuschränken, dass sie in angemessener Weise von Lina erwartet werden können. Anders ausgedrückt: Falls die genannten systembezogenen Aspekte von *JobAI* verstanden werden können, wären es Stakeholder wie Lina (nicht aber Stakeholder wie Fritz), von denen wir dieses Verständnis legitimerweise erwarten können.

Andere Überzeugungen von Linas Überzeugungssystem sind hingegen peripher. Hier ist ein Beispiel: *Das KI-System AmazingJob, ein ehemaliges Konkurrenzprodukt von JobAI, schätzte Bewerber:innen häufig falsch ein, weil sich AmazingJob maßgeblich auf das Geschlecht der Bewerber:innen stützte, um ihre Eignung für eine Stelle zu beurteilen.* Diese Überzeugung ist Teil von Linas Überzeugungssystem, insofern sie Verknüpfungen mit anderen ihrer Überzeugungen über *JobAI* aufweist (unter anderem, weil es sich um ein Konkurrenzprodukt handelt, das gewisse Ähnlichkeiten mit *JobAI* aufweist). In Anbetracht des Desideratums ist sie aber peripher: Zwar könnte es sein, dass *JobAI* aus ähnlichen Gründen versagt wie *AmazingJob*, doch handelt es sich um zwei verschiedene, voneinander unabhängige KI-Systeme. Sollte die Überzeugung falsch sein, erschwert dies die Erfüllung von Linas Desideratum, wenn überhaupt, nur geringfügig. Zudem erwarten wir von Lina *qua* ihrer Rolle nicht, dass sie sich detailliert mit Konkurrenzprodukten auskennt. Es wäre in dieser Hinsicht verzeihlich, wenn diese Überzeugung falsch sein sollte (oder wenn Lina keinerlei Überzeugung über *AmazingJob* besäße).

Linas Bedingung (1c) Bedingung (1c) verlangt einen kontextuell ausreichenden Rechtfertigungsgrad von Linas Überzeugungen. Der relevante Kontextfaktor sind die involvierten

Risiken bzw. Chancen. Wie viel steht auf dem Spiel, sollte Linas Desideratum (nicht) erfüllt werden? Im beschriebenen Beispiel steht viel auf dem Spiel. Zum einen in Hinblick auf moralische Werte: Sollte *JobAI* eine signifikante Anzahl von Bewerber:innen falsch einstufen, könnte dies zu einer ungerechten Behandlung dieser Bewerber:innen – und im vorliegenden Beispiel sogar zu systematischer Diskriminierung – führen. Zum anderen droht Linas Unternehmen ein ökonomischer Schaden, wenn sie den Fehler nicht lokalisieren und beheben kann. Ein fehlerhaftes, unzuverlässiges System verkauft sich schlecht und schädigt das Unternehmensimage. Das gefährdet auch Linas Stellung im Unternehmen.

Damit Lina instrumentell versteht, müssen ihre Überzeugungen daher gut gerechtfertigt sein. Das gilt vor allem für ihre zentralen Überzeugungen über die Ursache bzw. Beseitigung des Fehlers. Linas periphere Überzeugungen müssen hingegen weniger gut gerechtfertigt sein: Es wäre beispielsweise unproblematisch, wenn Linas oben genannte Überzeugung über *AmazingJob* nur schwach mit ihren übrigen Überzeugungen kohäriert (z. B., weil sie sonst nur wenig über das System weiß) oder wenn sie diese Überzeugung auf den wütenden Social-Media-Beitrag eines kürzlich gefeuerten Ex-Entwicklers von *AmazingJob* stützt, der sich an seinem ehemaligen Arbeitgeber rächen will. Denn diese Überzeugung hat keinen oder nur einen geringen Einfluss auf die Erfüllung des fraglichen Desideratums. Entsprechend ist weniger wichtig, dass sie (in Form von Rechtfertigung) angemessen mit der Wahrheit verbunden ist.

Linas Bedingung (2) Damit ist das Überzeugungssystem, das Lina für instrumentelles Verstehen benötigt, beschrieben. Bleibt noch Bedingung (2), welche die kontextuell erforderlichen verstehensbezogenen Fähigkeiten spezifiziert. Auch hier ist Linas Desideratum, den Fehler zu beheben, entscheidend. Diesbezüglich stechen (neben basalen Fähigkeiten des Erkennen-Clusters) insbesondere Fähigkeiten aus dem Intervenieren-Cluster hervor: Um instrumentell zu verstehen, muss Lina über die Fähigkeit verfügen, den Bias von *JobAI* zu identifizieren und zumindest mental zu simulieren, welche Änderungen nötig wären, um die Verzerrung zu beseitigen. Aufgrund ihrer Rolle und der daher zu erwartenden Expertise über das Empfehlungssystem ist es außerdem legitim, von Lina solche Fähigkeiten zu erwarten. Damit handelt es sich dabei um für Linas instrumentelles Verstehen erforderliche Fähigkeiten.

Fritz' instrumentelles Verstehen

Betrachten wir als nächstes das instrumentelle Verstehen von Personalmanager Fritz. Ich fasse mich kürzer als bei der Charakterisierung von Linas instrumentellem Verstehen, da viele der obigen Überlegungen auch für Fritz gelten. Fritz verwendet *JobAI*, um geeignete Bewerber:innen für eine vakante Stelle in seinem Unternehmen zu identifizieren. Wir können annehmen, dass sein Desideratum darin besteht, dass das System in Bezug auf die Beurteilung von Magdas Bewerbung vertrauenswürdig ist bzw. dass er dem System in dieser Hinsicht angemessen vertrauen kann. Die epistemische Erfüllung dieses Desideratums

erfordert, dass Fritz epistemischen Zugang dazu hat, ob dem so ist, bzw. ob das System Eigenschaften aufweist, welche die fragliche Entscheidung vertrauenswürdig machen. Von der Vertrauenswürdigkeit des Systems bzw. von Magdas Einstufung hängt einiges ab: Fritz sollte natürlich die bestmögliche Besetzung für die vakante Stelle auswählen. Außerdem muss er vermeiden, eine ungerechte Entscheidung zu fällen – etwa, indem er weibliche Bewerber:innen diskriminiert. Die Chancen und Risiken sind daher recht hoch.

Da Vertrauenswürdigkeit eines der zentralen Desiderata gegenüber KI-Systemen darstellt, ist die zugehörige Literatur umfangreich (eine aktuelle Übersicht bieten Durán & Pozzi, 2025). In Bezug auf Menschen wird Vertrauenswürdigkeit üblicherweise als ‚Zuverlässigkeit plus *X*‘ definiert. Kompetenz hinsichtlich der anvertrauten Aufgabe(n) ist demnach eine notwendige, aber nicht hinreichende Bedingung von Vertrauenswürdigkeit. Das zusätzlich erforderliche ‚*X*‘ wurde unterschiedlich ausbuchstabiert – etwa in Form von Wohlwollen, geeigneten Motivationen oder einer Verpflichtung (engl. *commitment*) aufseiten der Vertrauensnehmerin (ebd., Kap. 2, 4).

Doch was ist Vertrauenswürdigkeit im Kontext von KI? Für meine Zwecke ist nicht wichtig, für welche Definition man sich entscheidet. Nehmen wir exemplarisch an, dass sich beide Komponenten menschlicher Vertrauenswürdigkeit auch auf KI-Systeme übertragen lassen (Schmidt & Mann, o. D., Manuskript). Kompetenz bzw. Zuverlässigkeit von KI lässt sich durch geeignete Performance-Metriken (etwa Genauigkeit) ausdrücken. Als maschinentaugliches Analogon für das ‚*X*‘ wurde eine Übereinstimmung der ‚Ziele‘ des Systems mit jenen der vertrauenden Person vorgeschlagen (Schmidt, 2022). Wenn man diese Ziele als Dispositionen auffasst, setzt dieser Vorschlag keine genuinen Absichten, Wünsche o. ä. aufseiten des KI-Systems voraus (Schmidt & Mann, o. D.). Man könnte z. B. COMPAS das Ziel zuschreiben, Schwarze Angeklagte gegenüber weißen Personen zu benachteiligen, wenn das System disponiert ist, vor allem Schwarzen Personen fälschlicherweise eine hohe Rückfallwahrscheinlichkeit zu prognostizieren (vgl. Angwin et al., 2016; siehe Abschnitt 3.3.2) – unabhängig davon, ob dieses Ziel von den Entwickler:innen beabsichtigt wurde.

Übertragen wir diese Überlegungen auf Fritz’ instrumentelles Verstehen und beginnen mit Bedingung (1), die sein Überzeugungssystem näher bestimmt. Die Inhalte des Überzeugungssystems (Bedingung 1a) sind Propositionen über systembezogene Aspekte. Im vorliegenden Fall sind das insbesondere Aspekte von *JobAI* oder dem übergeordneten soziotechnischen System, insofern sie die Beurteilung von Bewerberin Magda betreffen. Dazu können z. B. auch Informationen über die vakante Stelle und deren Funktion im Unternehmen oder Eigenschaften (un)geeigneter Bewerber:innen zählen. Überzeugungen über solche systembezogenen Aspekte müssen *ultima facie* gerechtfertigt sein, wobei der erforderliche Rechtfertigungsgrad aufgrund der großen Chancen und Risiken recht hoch ist (Bedingung 1c).

Betrachten wir nun Fritz’ kontextuell zentrale Überzeugungen genauer (1b), die wenigstens annähernd wahr und besonders gut gerechtfertigt sein müssen. Für die Erfüllung seines Desideratums ist es wichtig, dass Fritz wahre und gut gerechtfertigte Überzeugungen

darüber besitzt, ob *JobAI* zuverlässig funktioniert (also z. B. eine hohe Richtig-positiv-Rate aufweist) und in Übereinstimmung mit Fritz' Zielen entscheidet. Diese Ziele können vielfältig sein: Fritz soll beispielsweise die vakante Stelle mit einer qualifizierten Person besetzen, zur Gleichstellung relevanter Personengruppen beitragen, eine datenschutzkonforme Verarbeitung der Bewerbungsunterlagen gewährleisten und den Auswahlprozess beschleunigen.

Da es ihm im vorliegenden Beispiel konkret um die Einstufung von Bewerberin Magda geht, braucht Fritz wahre und gut gerechtfertigte Überzeugungen über die Faktoren, die das System zu seiner Entscheidung bewogen haben, und welche Ziele im Hintergrund dieser Entscheidung stehen. Denn wenn *JobAI* seine Einschätzung z. B. vornehmlich auf Magdas Balletthobby stützt, hat das System keine kompetente Entscheidung getroffen und ist deswegen nicht vertrauenswürdig. Wenn eines der (impliziten) Ziele des Empfehlungssystems die Benachteiligung weiblicher Bewerberinnen oder das Ausspähen persönlicher Daten ist, erfüllt es die zweite Komponente (das ‚X‘) von Vertrauenswürdigkeit nicht. Diese Überzeugungen können angesichts seiner Rolle legitimerweise von Fritz erwartet werden: Als Personalmanager, der ein KI-System für die Auswahl von Bewerber:innen nutzt, können wir von ihm verlangen, dass er sich über die Zuverlässigkeit des Systems informiert (z. B., indem er Prüfsiegel oder Zertifikate heranzieht) und auf Endnutzer:innen wie ihn zugeschnittene Erklärungen konsultiert. Zwar würde auch ein detailliertes Verständnis der technischen Hintergründe zur Erfüllung seines Desideratums beitragen, doch wäre dies, *qua* seiner Rolle, zu viel verlangt.

Bleiben noch die Fähigkeiten, die Fritz braucht, um instrumentell zu verstehen (Bedingung 2). Er muss (zumindest privat) erklären können, warum *JobAI* Magda als ungeeignet eingestuft hat. Außerdem muss er in der Lage sein, die vom System herangezogenen Kriterien zu bewerten: Hatte das System gute Gründe dafür, Magdas Bewerbung lediglich drei Punkte zu geben? Im vorliegenden Beispiel ist das einfach – Magdas Ballettkenntnisse sollten, wenn überhaupt, nur eine marginale Rolle dafür spielen, ob sie die Stelle bekommt. Doch können wir von Fritz erwarten, dass er auch subtilere Proxies als problematisch einstufen kann: Aufgrund seiner Rolle als Personalmanager sollte er die zu besetzende Stelle kennen und bis zu einem gewissen Grad einschätzen können, welche Qualifikationen und Kenntnisse dafür von Relevanz sind.

Wenn instrumentelles Verstehen fehlt

Betrachten wir nun noch ein anderes Beispiel, um zu illustrieren, was passiert, wenn die Bedingungen instrumentellen Verstehens unerfüllt bleiben. Ziehen wir dafür noch einmal das *Röntgen*-Beispiel aus Abschnitt 9.2.1 heran:

Röntgen: Die Ärztin Marla verwendet ein KI-System, um auffällige Gewebestrukturen in Röntgenbildern zu identifizieren. Bei dem System handelt es sich um ein tiefes neuronales Netzwerk. Marla nutzt zusätzlich zu diesem System eine Erklärbarkeits-

methode (z. B. LIME), die Salienzkarten für die einzelnen Klassifikationen generiert. Das Röntgenbild der Lunge einer bestimmten Patientin wird vom KI-System als ‚maligne‘ klassifiziert. Die Salienzkarte hebt eine Stelle unten links im Röntgenbild hervor.

Nehmen wir an, dass die Erfüllung von Marlas Desideratum darin besteht, herauszufinden, ob die Klassifikation des Röntgenbilds akkurat bzw. zutreffend ist. Außerdem können wir stipulieren, dass viel auf dem Spiel steht: Wenn Marla Lungenkrebs diagnostiziert, obwohl die Patientin gesund ist, würde sie ihr unnötigerweise körperlich und psychisch äußerst belastende Behandlungen (z. B. eine Strahlen- oder Chemotherapie) verordnen. In diesem Beispiel ist unter anderem eine Überzeugung wie die folgende zentral: *Die Gewebeveränderungen, die sich unten links auf dem Röntgenbild befinden, haben die Klassifikation als ‚maligne‘ ausgelöst.* Was geschieht, wenn Marla eine der Bedingungen instrumentellen Verstehens nicht erfüllt?

Laut Bedingung (1a) braucht Marla ein Überzeugungssystem über systembezogene Aspekte. Die Notwendigkeit dieser Bedingung scheint mir eindeutig – wenn Marla keinerlei Überzeugungen über Aspekte hat, die das medizinische Klassifikationssystem direkt oder indirekt betreffen, kann sie schwerlich beurteilen, ob die Klassifikation des Systems zutreffend ist. Sie hat dann beispielsweise keine Meinung dazu, was das System zu welcher Klassifikation veranlasst hat, oder keine Expertise in der Anwendungsdomäne des Systems.

Bedingung (1b) verlangt, dass Marlas Überzeugungen über solche systembezogenen Aspekte moderat faktiv sind. Es ist demnach unproblematisch, wenn Marla einzelne falsche Überzeugungen über das Klassifikationssystem besitzt, solange diese peripher sind. Zentrale Überzeugungen müssen hingegen (näherungsweise) wahr sein. Wir haben oben bereits eine kontextuell zentrale Überzeugung identifiziert. Nehmen wir also an, dass Marla es für falsch hält, dass das unten links sichtbare Gewebe zur Klassifikation als ‚maligne‘ geführt hat. Vielleicht glaubt sie, dass ein anderer Bildbereich dafür verantwortlich war, oder dass das Klassifikationssystem eine problematische Verzerrung aufweist und die ethnische Zugehörigkeit der Patientin zur Klassifikation geführt hat. In diesen Szenarien fehlt ihr das Verständnis, das ihr die Erfüllung ihres Desideratums ermöglichen würde: Sie kann nicht mithilfe ihres Verstehens beurteilen, ob die Klassifikation des Systems korrekt ist.

Subbedingung (1c) erfordert eine ausreichende Rechtfertigung von Marlas Überzeugungen. Stellen wir uns vor, dass Marla ihre (zufälligerweise wahren) Überzeugungen auf das Video eines technikassenden Youtubers oder auf das Onkologie-Buch eines bekannten Scharlatans stützt. Dann fehlt ihr ebenfalls jenes Verstehen, das ihr die adäquate Beurteilung der Klassifikation des Systems ermöglichen würde: In einem solchen Fall besteht keine angemessene Verknüpfung zwischen ihren Überzeugungen und der Wahrheit. Schließlich ist es bloßer Zufall, dass Marla die Ursache für die Klassifikation identifiziert hat.

Außerdem verlangt Bedingung (1), dass Marla ein adäquat verknüpftes Überzeugungssystem (anstatt einzelne, isolierte Überzeugungen) besitzt. Wir können uns vorstellen, dass Marla alle zentralen Überzeugungen besitzt und dass diese wahr und fundamentalistisch

gut gerechtfertigt sind, ohne aber miteinander verbunden zu sein – etwa, wenn Marla gerechtfertigterweise und wahrheitsgemäß glaubt, dass der Bereich unten links im Bild die Klassifikation ausgelöst hat, und dass das Röntgenbild als ‚maligne‘ klassifiziert wurde, und dass das Bild die Lunge von Patientin X zeigt usw., ohne dass diese Überzeugungen angemessen miteinander verknüpft sind. Abermals scheint klar zu sein, dass Marla auf dieser Grundlage nicht beurteilen kann, ob die Klassifikation des Systems korrekt ist – schließlich begreift sie nicht, wie relevante Merkmale der Eingabedaten und die Klassifikation miteinander zusammenhängen.

Nehmen wir abschließend an, dass Bedingung (2) unerfüllt bleibt, also dass Marla kontextuell relevante Fähigkeiten fehlen. Das wäre beispielsweise der Fall, wenn Marla die Klassifikation nicht einmal privat erklären kann, wenn sie die Bedeutung der Salienzkarte nicht erkennt und natürlich, wenn sie unfähig ist zu beurteilen, ob die Klassifikation angemessen ist oder nicht. In Ermangelung dieser Fähigkeiten scheint Marla abermals nicht jenes Verstehen zu besitzen, das ihr die Erfüllung ihres Desideratums ermöglicht.

Zusammengefasst: Während die Diskussion zu Lina und Fritz aus dem *Bewerbung*-Beispiel nahelegt, dass die vorgeschlagenen Bedingungen hinreichend für instrumentelles Verstehen sind, spricht die Betrachtung von *Röntgen* dafür, dass sie auch notwendig sind. Bleibt eine der Bedingungen unerfüllt, fehlt jenes Verständnis, das dem fraglichen Stakeholder die Erfüllung seines Desideratums ermöglicht – er verfehlt instrumentelles Verstehen.

Ich behaupte nicht, dass instrumentelles Verstehen notwendig für die (epistemische) Erfüllung gesellschaftlicher Desiderata ist. Manchmal genügt dafür auch atomistisches ‚instrumentelles Wissen‘ über den Status des fraglichen Desideratums, das z. B. durch das Zeugnis einer Expertin erworben werden kann (siehe Abschnitt 6.3). Diese Option motiviert aber nicht diese Arbeit. Das Black-Box-Problem ergibt sich erst, weil Stakeholdern der Status ihrer Desiderata nicht ‚auf dem Silbertablett‘ serviert wird. Mein Explikat betrifft daher jene Fälle, in denen ein Stakeholder selbst ein gewisses Maß epistemischer Arbeit investieren muss, um seine Desiderata zu erfüllen – etwa, indem er XAI-Erklärungen konsultiert.

11.1.3 Zum Verhältnis von Wissen und (instrumentellem) Verstehen

Nun, da instrumentelles Verstehen definiert ist, können wir uns einer Frage zuwenden, die im Hintergrund vieler der in Kapitel 5 bis 10 verhandelten Debatten steht: Ist Wissen notwendig und/oder hinreichend für (instrumentelles) Verstehen? Diese Fragestellung ist für mich nicht entscheidend: In dieser Arbeit geht es mir um die Charakterisierung eines spezifischen epistemischen Zustands, der einen konkreten Zweck in XAI-Kontexten erfüllen soll. Sollte sich herausstellen, dass dieser epistemische Zustand eine Wissensform oder reduzibel auf Wissen ist, bin ich damit ebenso einverstanden wie mit dem gegenteiligen Ergebnis. Aufgrund der wichtigen Stellung, die das Verhältnis von Verstehen und Wissen in den hier relevanten philosophischen Debatten einnimmt, möchte ich dennoch kurz darauf eingehen.

Erwartungsgemäß gibt es für jede der erwähnten Positionen – dass Wissen notwendig und/oder hinreichend für (explanatorisches und/oder objektuales) Verstehen ist bzw. dass dem nicht so ist – sowohl engagierte Befürworter:innen als auch entschiedene Gegner:innen.¹²⁶ Die meisten Argumentationslinien verlaufen dabei entlang der Eigenschaften von Verstehen, die ich im Verlauf dieser Arbeit diskutiert habe. Kehren wir also noch einmal zu diesen Debatten zurück.

In Kapitel 5 habe ich argumentiert, dass Verstehen *simpliciter* gradierbar ist, während sich in Bezug auf instrumentelles Verstehen keine Grade unterscheiden lassen. Der Standardauffassung zufolge ist Wissen nicht gradierbar – entweder, man weiß, dass *p* oder man weiß es nicht. In dieser Hinsicht ist instrumentelles Verstehen näher an Wissen als Verstehen *simpliciter*. Allerdings ist instrumentelles Verstehen deswegen nicht gradierbar, weil es in einem bestimmten, kontextsensitiven Verstehensgrad besteht. Die fehlende Gradierbarkeit von Wissen rührt hingegen nicht daher, dass es auf einen spezifischen Wissensgrad zurückgeführt werden kann. Das spricht gegen die Reduzierbarkeit von (instrumentellem) Verstehen auf Wissen.

In Kapitel 6 habe ich argumentiert, dass die Objekte von Verstehen *simpliciter* und instrumentellem Verstehen größere, zusammenhängende Informationsmengen sind, die eine gewisse Komplexität aufweisen. Im Kontrast dazu ist das Objekt propositionalen Wissens zumindest in paradigmatischen Fällen eine individuelle Proposition. In dieser Hinsicht zeichnet sich also ein weiterer Unterschied zwischen Verstehen und Wissen ab. Wenn man jedoch annimmt, dass holistisches Verstehen einer größeren Menge atomistischen Wissens entspricht, könnte Verstehen in dieser Hinsicht als auf Wissen reduzierbar aufgefasst werden.

Wie ich in Kapitel 7 aufgezeigt habe, versuchen einige Philosoph:innen, anhand der jeweils involvierten epistemischen Einstellung einen Keil zwischen Verstehen und Wissen zu treiben. Meine Diskussion spricht hingegen dafür, dass Verstehen *simpliciter* und instrumentelles Verstehen – genau wie Wissen – Überzeugung erfordern. Wenn das stimmt, besteht in dieser Hinsicht keine Diskrepanz zwischen Verstehen und Wissen.

Viele Autor:innen haben sich in Bezug auf Verstehen gegen einen strengen Faktivismus ausgesprochen. Wissen ist in seiner paradigmatischen Form hingegen faktiv: Man kann nicht wissen, dass *p*, ohne dass es der Fall ist, dass *p*. Diese Diskrepanz zwischen Wissen

¹²⁶ Zu den Autor:innen, die Verstehen als eine bestimmte Form von Wissen auffassen – die also annehmen, dass Wissen notwendig, aber nicht hinreichend für Verstehen ist –, zählen Cooper, 1994; Gijssbers, 2013; Greco, 2014; Grimm, 2006, 2014; Kelp, 2015, 2017, 2021, und Strevens, 2013. Andere nehmen an, dass Verstehen auf Wissen reduziert werden kann – dass Wissen also notwendig und hinreichend für Verstehen ist (Brogaard, 2005; P. Lipton, 1990, 2009; Riaz, 2015; Sliwa, 2015). Kareem Khalifa zufolge ist zumindest maximales Verstehen auf Wissen reduzierbar (Khalifa, 2017). Viele andere, insbesondere erkenntnistheoretische Beiträge argumentieren hingegen, dass Verstehen nicht einmal eine Form des Wissens ist, weil Wissen nicht notwendig für Verstehen ist (Baumberger, 2011; Dellsén, 2017; Elgin, 2007, 2017; Hills, 2016; Kvanvig, 2003; Moravcsik, 1979; Newman, 2017a; Pritchard, 2008, 2010, 2014; Wilkenfeld, 2017b; Zagzebski, 2001). Miloud Belkoniene (2023) schlägt umgekehrt vor, Wissen als bestimmten Typ von Verstehen aufzufassen. Eine Übersicht über die zugehörigen Subdebatten, die ich nachfolgend nur skizzieren kann, bieten Christoph Baumberger et al. (2017) und Michael Hannon (2021).

und Verstehen lässt sich darauf zurückführen, dass ersteres üblicherweise atomistisch und letzteres holistisch ist: In Kapitel 8 habe ich dargelegt, dass die Faktizität von Verstehen häufig davon abhängig gemacht wird, ob es sich um propositionales, interrogatives oder objektuales Verstehen handelt. Wenn die von mir vertretene Position korrekt ist, sind hingegen alle drei Formen von Verstehen moderat faktiv, weil sie alle ein komplexes Verstehensobjekt betreffen. Diese Komplexität macht wiederum einzelne falsche Überzeugungen mit (instrumentellem) Verstehen kompatibel und spricht dagegen, dass Verstehen auf Wissen reduziert werden kann.

Während die meisten Autor:innen einen engen Zusammenhang zwischen Verstehen und Fähigkeiten annehmen, ist umstritten, wie umfangreich die in Wissen involvierten Fähigkeiten sind. Sollte meine Argumentation in Kapitel 9 zutreffend sein, würde zumindest instrumentelles Verstehen (teils anspruchsvolle) Fähigkeiten erfordern. Es ist *prima facie* attraktiv, denselben Anspruch nicht auch an Wissen zu stellen. Zum einen ist es intuitiv plausibel, dass wir diverse Dinge wissen, über die wir beispielsweise keine nennenswerten Schlüsse ziehen können – der alltagssprachliche Wissensbegriff scheint keine umfangreichen Fähigkeiten vorauszusetzen. Zum anderen bleibt Wissen dadurch eine häufige, vergleichsweise einfach zu erlangende epistemische Errungenschaft. Würde man verstehensbezogene Fähigkeiten zu einer notwendigen Bedingung für Wissen machen, würde Wissen zu einem deutlich rarerem epistemischen Gut. Diese Überlegungen legen nahe, dass Wissen zumindest nicht hinreichend für Verstehen ist.

In Kapitel 10 habe ich argumentiert, dass Verstehen – wie Wissen – Rechtfertigung erfordert. Allerdings habe ich auch vorgeschlagen, dass sich für Verstehen eine internalistische, hybride Rechtfertigungstheorie anbietet, weil diese Art der Rechtfertigung eng mit dem für Verstehen charakteristischen Begreifen von Zusammenhängen verknüpft ist. Zwar wurden viele der existierenden hybriden Rechtfertigungstheorien ursprünglich für Wissen vorgeschlagen, doch könnte es sein, dass das kohärentistische Element der Rechtfertigung für Wissen nicht erforderlich ist (diese Position vertritt etwa Schmidt, 2026, Fn. 13 auf S. 230). Wenn dem so ist, würden sich Verstehen und Wissen in dieser Hinsicht unterscheiden.

Eine zugehörige Frage betrifft die Kompatibilität von Verstehen mit epistemischem Zufall. Epistemischer Zufall liegt vor, wenn eine Person eine gerechtfertigte, wahre Überzeugung ausbildet, obwohl diese leicht hätte falsch sein können. Seit Edmund Gettier (1963) wird üblicherweise davon ausgegangen, dass gerechtfertigte wahre Überzeugung nicht hinreichend für Wissen ist, weil bestimmte Formen des epistemischen Zufalls diesen Status untergraben. Ob dasselbe auch für Verstehen *simpliciter* gilt, ist umstritten.¹²⁷

¹²⁷Unter anderem Greco, 2014; Grimm, 2006, 2014; Kelp, 2017, 2021; Riaz, 2015, und Sliwa, 2015, vertreten die Auffassung, dass Verstehen in gleicher Form wie Wissen durch epistemischen Zufall untergraben wird. Andere argumentieren, dass Verstehen mit Arten von epistemischem Zufall kompatibel ist, die Wissen verhindern (etwa Baumberger, 2011; Belkoniene, 2023, Kap. 4.5; Hills, 2016; Kvanvig, 2003, Kap. 8; Newman, 2017a; Pritchard, 2008, 2010). Berit Brogaard (2005) stellt eine Anti-Zufalls-Bedingung gleichermaßen für Wissen wie Verstehen in Frage.

Ich neige dazu anzunehmen, dass weder Verstehen *simpliciter* noch instrumentelles Verstehen mit den relevanten Formen epistemischen Zufalls vereinbar sind. Hier ist ein Grund dafür, eine solche Inkompatibilität zumindest in Bezug auf instrumentelles Verstehen anzunehmen: Eine wichtige und plausible Forderung gegenüber XAI-Forscher:innen lautet, dass diese zuverlässige und aussagekräftige Erklärbarkeitsmethoden entwickeln sollen. In XAI-Kontexten wollen wir sicherstellen, dass die entsprechenden Verstehensüberzeugungen über relevante systembezogene Aspekte *aufgrund* solch geeigneter Verstehensmöglichkeiten zustande kommen, und dass Desiderata *wegen* dieser Ansätze (epistemisch) erfüllt werden. In Fällen, in denen epistemischer Zufall im Spiel ist, besteht diese Beziehung nicht. Die entsprechenden Überzeugungen kommen nicht aufgrund eines geeigneten Verstehensmöglichkeiten, sondern dank glücklicher Umstände zustande und hätten daher leicht falsch sein können. Intuitiv ist ein solcher Fall ebenso problematisch wie wenn es um den Erwerb von Wissen ginge. Wenn dem so ist, gleichen sich Wissen und (instrumentelles) Verstehen in dieser Hinsicht.

Eine letzter möglicher Unterschied zwischen Verstehen und Wissen ist in der bisherigen Diskussion noch nicht aufgekommen: die testimoniale Übertragbarkeit. Der Standardauffassung zufolge ist propositionales Wissen testimonial leicht übertragbar: Wenn eine Expertin weiß, dass *p*, kann eine Hörerin auf Grundlage ihrer Aussage, dass *p*, zu der gerechtfertigten, wahren und nicht durch epistemischen Zufall beeinträchtigten Überzeugung (und damit Wissen) kommen, dass *p*. Einige Autor:innen haben argumentiert, dass Verstehen im Gegensatz dazu nicht (oder nur schwerlich) testimonial erworben werden kann (etwa Cooper, 1994, S. 5–7; Hills, 2016, S. 670–671; Zagzebski, 2009, S. 145–146). Häufig wird als Grund angeführt, dass Verstehen eine besondere Einsicht verlange (das vielfach erwähnte Begreifen von Zusammenhängen). Diese Einsicht erfordere wiederum einen gewissen Aufwand aufseiten des verstehenden Subjekts. Das Anhören einer zuverlässigen Sprecherin gewährleiste allerdings nicht, dass die Hörerin diesen Aufwand betreibt und die relevanten Zusammenhänge wirklich begreift (eine ausführliche Diskussion und Kritik an dieser Argumentation bietet Malfatti, 2023, Kap. 6).

Ich sympathisiere mit einer Position, die in ähnlicher Form von Kenneth Boyd (2017), Stephen Grimm (2020), Federica Malfatti (2023) und Paulina Sliwa (2015) vorgebracht wurde: Sowohl Wissen als auch Verstehen sind in manchen Fällen leicht testimonial übertragbar, in anderen nicht. Der relevante Faktor ist nicht das jeweils zu übertragende epistemische Gut, sondern kontextuelle Faktoren. Wie Grimm es ausdrückt: Damit der testimoniale Transfer von Wissen und Verstehen erfolgreich ist, müssen geeignete ‚Aufnahmebedingungen‘ (engl. *uptake conditions*; z. B. relevantes Hintergrundwissen oder -verstehen) bei der Hörerin vorhanden sein (Grimm, 2020, S. 135). Fehlen diese Aufnahmebedingungen, ist der Transfer erschwert oder gar unmöglich. Dieser Argumentation zufolge markiert die testimoniale Übertragbarkeit von Verstehen bzw. Wissen keinen relevanten Unterschied zwischen beiden epistemischen Gütern.

Zusammengefasst sprechen die Gradierbarkeit, die Komplexität der Objekte, die moderierte Faktizität, die Notwendigkeit von Fähigkeiten und die kohärentistische Rechtfertigung

von Verstehen dafür, dass Wissen nicht hinreichend und/oder nicht notwendig für (instrumentelles) Verstehen ist. Natürlich kann man an verschiedenen Stellen dieser knappen Gegenüberstellung einhaken und entweder Verstehen näher an Wissen rücken (etwa, indem man die Notwendigkeit von Fähigkeiten verneint) und/oder umgekehrt (z. B., indem man Wissen als gradierbar auffasst). Dies könnte dazu führen, dass auch (instrumentelles) Verstehen als eine Form des Wissens oder als auf Wissen reduzierbar gelten muss. Wie eingangs geschildert, habe ich nichts gegen dieses Ergebnis einzuwenden. Insgesamt betrachtet spricht derzeit aber mehr dafür, dass es sich bei Verstehen *simpliciter* und instrumentellem Verstehen um einen epistemischen Zustand handelt, der sich in verschiedenen Hinsichten von Wissen unterscheidet.

11.2 Evaluation des Explikats

In diesem Abschnitt schließe ich meine Explikation instrumentellen Verstehens mit dem vierten und letzten Schritt des Explikations-Rezepts (Abschnitt 4.1.3) ab: Ich überprüfe, ob das Explikat die in Abschnitt 4.3 formulierten Adäquatheitsbedingungen erfüllt. In Abschnitt 11.2.1 evaluiere ich zunächst die erforderliche Ähnlichkeit zwischen Explikat und Explikandum (Schritt 4a). Anschließend überprüfe ich in Abschnitt 11.2.2, ob das Explikat den gewünschten Nützlichkeitskriterien gerecht wird (4b).

11.2.1 Evaluation der Ähnlichkeit

In Abschnitt 4.3.1 habe ich für ein semantisches Ähnlichkeitskriterium plädiert, um die begriffliche Kontinuität zwischen Explikat und Explikandum sicherzustellen bzw. um zu verhindern, dass meine Explikation einfach ‚das Thema wechselt‘. Motiviert habe ich dieses Vorgehen dadurch, dass ich eine bestimmte Form des Verstehens explizieren möchte, die innerhalb der XAI-Literatur durch Verweise auf Verstehen *simpliciter* charakterisiert wird. Ein wichtiges Ziel der vorangegangenen Auseinandersetzung mit philosophischen Debatten zu Verstehen war daher, für instrumentelles Verstehen relevante Merkmale von Verstehen *simpliciter* herauszuarbeiten und dadurch zu gewährleisten, dass es sich bei instrumentellem Verstehen wirklich um *Verstehen* handelt. Dabei wurde deutlich, dass Explikat und Explikandum zahlreiche Parallelen aufweisen: Sowohl instrumentelles Verstehen als auch Verstehen *simpliciter* erfordern Überzeugung sowie Rechtfertigung, haben komplexe Objekte, sind moderat faktiv und gehen mit Fähigkeiten einher.

Ein weiterer Indikator für die große semantische Ähnlichkeit zwischen Verstehen *simpliciter* und instrumentellem Verstehen ist die Übereinstimmung meines Vorschlags mit bestehenden Verstehensdefinitionen. Wie auch Christoph Baumbergers Definition objektualen Verstehens mithilfe einer Theorie (Baumberger, 2019, S. 376) weist meine Definition eine Verpflichtungs-, Richtigkeits-, Begreifens- und Rechtfertigungskomponente auf. Verpflichtung bestimme ich jedoch anders als Baumberger nicht in Form von Akzeptanz (ebd., Kap. 4.4), sondern Überzeugung. Der wichtigste Unterschied besteht aber in der Rolle

von Verstehensmöglichkeiten. Die Faktizität von Verstehen hängt bei Baumberger allein von der Faktizität der begriffenen Theorie ab. In den Kapiteln 7 und 8 wurde allerdings deutlich, dass die Wahrheit von Verstehensüberzeugungen prinzipiell unabhängig von der Akkuratheit einer Theorie, Erklärung oder sonstigen Repräsentation ist. Angesichts dessen ziehe ich es vor, die Faktizität von Verstehen an die Wahrheit der involvierten Überzeugungen zu knüpfen.

Ein weiterer Vorschlag, der deutliche Ähnlichkeiten zu meiner Definition aufweist, stammt von Laura Callahan. Auf Grundlage der Verstehensliteratur formuliert sie eine ‚Standardauffassung von Verstehen‘, derzufolge Verstehen kognitive Fähigkeiten sowie ein Cluster an Überzeugungen erfordert, dessen propositionale Inhalte angemessen miteinander verbunden sind und dessen Verknüpfungen vom verstehenden Subjekt begriffen werden (Callahan, 2018, S. 442).¹²⁸ Anders als ich stellt Callahan kein Faktizitätskriterium auf, allerdings scheint sie vorauszusetzen, dass die geglaubten Propositionen wahr sind (siehe z. B. ebd., S. 441, 450–452). Wenn das stimmt, wäre ihr Vorschlag dem streng faktiven Lager zuzuordnen, während ich für einen moderat faktiven Ansatz plädiert habe.

Schließlich ähnelt meine Definition den vier notwendigen Bedingungen für explanatorisches Verstehen, die Eva Schmidt aufstellt. Auch Schmidts Vorschlag verlangt, dass das epistemische Subjekt Überzeugungen hinsichtlich einer größeren Menge an Propositionen besitzt (auch wenn Schmidt sich im Gegensatz zu mir vorbehält, dass Akzeptanz eine Alternative zu Überzeugung darstellen könnte). Wie Baumberger und Callahan formuliert Schmidt ebenfalls eine Begreifensbedingung und deutet an, dass diese in Form verstehensbezogener Fähigkeiten ausbuchstabiert werden könnte (Schmidt, 2026, S. 230). Jedoch legt sich Schmidt, anders als ich, ausdrücklich nicht fest, ob Verstehen streng oder nur moderat faktiv ist. In der moderat faktiven Variante von Schmidts Vorschlag müssen die zentralen Überzeugungen wahr oder zumindest annähernd wahr sein – diesen Aspekt habe ich von Schmidt übernommen. Schließlich vertritt sie wie ich die Auffassung, dass sich für Verstehen eine hybride Theorie der Rechtfertigung am besten eignet.

Außerdem weist meine Definition offensichtlich Parallelen zu Definitionen von Verstehen auf, die Kontextfaktoren berücksichtigen (Bachmann, 2020; Baumberger, 2019; Carter, 2014; de Regt & Dieks, 2005; Egler, 2021; Kelp, 2015, 2017; Schmidt, 2026; Søndergaard, 2023; vgl. auch Páez, 2019). Im Gegensatz zu Relativist:innen, Kontrastivist:innen oder Kontextualist:innen vertrete ich hinsichtlich der Kontextsensitivität von Verstehen einen invariantistischen Ansatz.¹²⁹ Entscheidend ist also nicht der Bewertungskontext, die kon-

¹²⁸ Callahan erweitert diese Standardauffassung im weiteren Verlauf ihres Aufsatzes um adäquate affektive und motivationale Reaktionen, um bestimmten Intuitionen im Kontext moralischen Verstehens gerecht zu werden (Callahan, 2018, S. 451–452; siehe dazu Fußnote 108).

¹²⁹ Die Unterscheidungen stammen aus der Wissensdebatte, lassen sich aber auch auf Verstehen übertragen. Der *kontextualistischen* Lesart zufolge hängt die Wahrheit von Wissenszuschreibungen vom Sprecherkontext ab – z. B. davon, ob für die Sprecherin salient ist, dass für das potenziell wissende Subjekt viel auf dem Spiel steht. *Invariantist:innen* gehen hingegen davon aus, dass Wissenszuschreibungen von nicht-epistemischen Eigenschaften des Subjekts, dem Wissen (nicht) zugeschrieben wird, abhängen – zum Beispiel davon, wie viel für das Subjekt objektiv betrachtet auf dem Spiel steht. Eine invariantistische Position bezüglich

textuell salienten Alternativen oder der Kontext der Sprecherin, die einem epistemischen Subjekt Verstehen zuschreibt. Stattdessen kommt es auf die Eigenschaften des Kontexts an, in dem sich das epistemische Subjekt befindet – im Fall instrumentellen Verstehens also auf das Desideratum des Stakeholders, seine Rolle und die kontextuellen Risiken bzw. Chancen.

Insgesamt betrachtet besteht also eine große semantische Ähnlichkeit zwischen instrumentellem Verstehen und Verstehen *simpliciter*. Das überrascht nicht, handelt es sich bei instrumentellem Verstehen doch um eine bestimmte, auf einen spezifischen Zweck ausgerichtete Verstehensform. Mein Ziel war nicht, mich radikal von der etablierten Verstehensdebatte zu distanzieren. Vielmehr wollte ich einen auf die Erfordernisse von XAI-Kontexten zugeschnittenen Verstehensbegriff erarbeiten. Daraus ergibt sich auch der Mehrwert instrumentellen Verstehens gegenüber Verstehen *simpliciter*. Letzteres ist ein breiter Sammelbegriff für verschiedene kognitive und teils auch affektive Zustände oder Prozesse in alltäglichen oder wissenschaftlichen Kontexten (siehe Abschnitt 4.2.1). Instrumentelles Verstehen ist hingegen XAI-spezifisch und macht explizit, wie Desiderata mithilfe von Verstehen erfüllt werden können und welche für XAI relevanten Kontextfaktoren sich in welcher Form auf die einzelnen Verstehensbedingungen auswirken. Entscheidend ist auch der Ansatz, instrumentelles Verstehen so anspruchslos wie möglich zu machen. Insbesondere durch die Lockerung des Faktizitätskriteriums wird instrumentelles Verstehen für eine Vielzahl von Stakeholdern, einschließlich Nichtfachleuten, erreichbar. Bei alltäglichem oder wissenschaftlichem Verstehen gelten dagegen höhere Ansprüche an Faktizität, Fähigkeiten oder Vollständigkeit, oder wir erachten einen höheren Verstehensgrad zumindest als wünschenswert. Aufgrund seiner strengen Zweckgerichtetheit stellen sich diese Anforderungen für instrumentelles Verstehen nicht gleichermaßen.

Aus der Zweckbindung instrumentellen Verstehens ergeben sich weitere Unterschiede zu Verstehen *simpliciter*. Instrumentelles Verstehen ist nicht gradierbar: Entweder eine Instanz instrumentellen Verstehens ermöglicht die Erfüllung des relevanten Desideratums – oder eben nicht. Außerdem ist die Zentralität der in instrumentellem Verstehen involvierten Überzeugungen sowie die Notwendigkeit spezifischer verstehensbezogener Fähigkeiten aufgrund des zweckgebundenen Charakters kontextabhängig. Bei nicht an einen bestimmten Zweck gekoppelten Formen von Verstehen *simpliciter* ist es hingegen plausibel, dass Fähigkeiten bzw. wahre Überzeugungen existieren, die unabhängig vom

Verstehen vertritt Eva Schmidt (2026), einen kontextualistischen Vorschlag macht Marcus Bachmann (2020). Daniel Wilkenfeld (2015) lässt bewusst offen, ob sein Vorschlag zur Quasi-Faktizität objektualen Verstehens (vgl. Abschnitt 8.3.3) invariantistisch oder kontextualistisch aufgefasst werden soll. Der *Relativismus* nimmt an, dass die Standards des Kontexts, in denen Wissenszuschreibungen bewertet werden, entscheidend für die Wahrheit von Äußerungen wie ‚S weiß, dass p ‘ sind – und nicht jene des Kontexts des epistemischen Subjekts (Invariantismus) oder jene, in denen sie geäußert werden (Kontextualismus). Für eine Anwendung von John MacFarlanes Relativismus auf explanatorisches Verstehen siehe Adam Carter (2014) und Andreas Søndergaard (2023). Der *Kontrastivismus* charakterisiert Wissen als dreistellige Relation zwischen einem Subjekt, einem Faktum p und einer Alternative zu p . Der Kontrast kann je nach Kontext wechseln. Für eine kontrastivistische Position in Bezug auf explanatorisches Verstehen siehe Miguel Egler (2021).

Kontext notwendig für dieses Verstehen sind.

11.2.2 Evaluation der Nützlichkeit

Die Nützlichkeit meines Explikats bemisst sich daran, ob es im angepeilten theoretischen Rahmen die theoretischen Zwecke des Explikandums erfüllt und ausreichende Erklärungskraft besitzt, ob es klarer und exakter als das Explikandum sowie konsistent und fruchtbar ist.

Evaluation der Erfüllung der theoretischen Zwecke und der Erklärungskraft

Mein Begriff instrumentellen Verstehens soll mehreren theoretischen Zwecken im Kontext philosophischer, insbesondere erkenntnistheoretischer Literatur dienen. Einer davon ist, *grundsätzlich* zu erklären, inwiefern eine bestimmte Form des Verstehens die Mediatorfunktion zwischen Erklärbarkeitsansätzen (bzw. anderen Verstehensermöglichern) und gesellschaftlichen Desiderata erfüllen kann. Mein Explikat stellt einen wichtigen Schritt hin zur Klärung dieser Mediatorrolle dar: Instrumentelles Verstehen ist gegeben, wenn ein Stakeholder ein gerechtfertigtes, moderat faktives Überzeugungssystem über relevante systembezogene Aspekte besitzt sowie über relevante verstehensbezogene Fähigkeiten verfügt. Dieser epistemische Zustand stellt den intuitiv plausiblen Zusammenhang zu gesellschaftlichen Desiderata her: Ein geeignetes Überzeugungssystem und daraus resultierende Fähigkeiten erlauben es einem Stakeholder, sein Desideratum epistemisch und mitunter auch substanziell zu erfüllen.

Mein Explikat nimmt hingegen keinen Bezug auf XAI-Methoden oder andere Verstehensermöglicher. Einer der Aspekte, zwischen denen (instrumentelles) Verstehen medieren soll, wird von meiner Definition also nicht ausdrücklich berücksichtigt. Der Grund ist, dass ich instrumentelles Verstehen als epistemischen Zustand charakterisiert habe, der zunächst einmal unabhängig von einem spezifischen Vehikel für den Erwerb dieses Zustands ist. In Kapitel 12 werde ich allerdings wünschenswerte Eigenschaften von Verstehensermöglichern in XAI-Kontexten aufzeigen, die den Erwerb instrumentellen Verstehens begünstigen.

Weiterhin soll mein Explikat erklären können, warum Verstehen *in einem konkreten Kontext* die epistemische (und mitunter auch substanzielle) Erfüllung von Desiderata ermöglicht. Meine Definition trägt daher ausdrücklich dem Einfluss kontextueller Faktoren Rechnung. Meinem Vorschlag zufolge sind es das Desideratum des Stakeholders, ggf. ergänzt um seine Rolle, sowie die kontextuellen Risiken bzw. Chancen, die sich auf die Bedingungen instrumentellen Verstehens auswirken. Die Diskussion der Beispiele in Abschnitt 11.1.2 spricht dafür, dass mein Begriff instrumentellen Verstehens den Zusammenhang zwischen Verstehen und Desiderata in konkreten Kontexten erklärt, auch wenn ich dies nicht für alle Kontexte zeigen kann.

Anhand des *Bewerbung*-Beispiels habe ich exemplarisch ausbuchstabiert, worin instrumentelles Verstehen in unterschiedlichen Kontexten besteht. Anhand von *Röntgen* habe ich außerdem aufgezeigt, dass die Abwesenheit von instrumentellem Verstehen erklären kann, warum ein Desideratum *nicht* erfüllt wird. Sofern sich dieses Vorgehen auf weitere Kontexte übertragen lässt, kann meine Definition dabei helfen zu erklären, warum ein Desideratum in einem konkreten Kontext mithilfe von Verstehen erfüllt bzw. nicht erfüllt wird. Damit besitzt mein Explikat auch die gewünschte Erklärungskraft.

Evaluation der Klarheit, Exaktheit und Konsistenz

Die vorgeschlagene Definition instrumentellen Verstehens ist zweifellos klarer und exakter als die nicht näher spezifizierten Verweise auf Verstehen, die in der XAI-Literatur üblicherweise erfolgen. Nichtsdestotrotz bringt die starke Kontextabhängigkeit von instrumentellem Verstehen eine gewisse Vagheit mit sich. Es wäre wünschenswert, diese Vagheit weiter zu reduzieren – etwa, indem man feinkörniger zwischen verschiedenen Kontexten bzw. relevanten Kontextfaktoren unterscheidet. Mein Explikat könnte dadurch zum Explikandum einer konsekutiven Explikation werden. Allerdings ist davon auszugehen, dass die Vielfalt der XAI-Kontexte eine abschließende Untersuchung impraktikabel macht und eine gewisse Vagheit nicht eliminierbar ist.

Was das Adäquatheitskriterium der Konsistenz betrifft, weist meine Definition – soweit ich sehen kann – keine Widersprüche auf.

Evaluation der Fruchtbarkeit

Das letzte Adäquatheitskriterium ist die Fruchtbarkeit des Explikats. Die Erfüllung der oben genannten theoretischen Zwecke – insbesondere die Klärung der Mediatorrolle von Verstehen in XAI-Kontexten – trägt auch zur Fruchtbarkeit des Explikats bei: Ein substanzieller Verstehensbegriff ist fruchtbar für die weitere philosophische und interdisziplinäre Auseinandersetzung mit dem Zusammenhang von Erklärbarkeit und gesellschaftlichen Desiderata. Darüber hinaus habe ich in Abschnitt 4.3.2 folgende Aspekte der Fruchtbarkeit herausgestellt: das Ermöglichen universaler, aber auch partikularer Aussagen hinsichtlich unterschiedlicher Kontexte; die Möglichkeit, empirische und informatische XAI-Forschung erkenntnistheoretisch zu informieren sowie ein großer Geltungsbereich des Explikats.

Formulierung universaler und partikularer Aussagen Carnap zufolge sind insbesondere universale Aussagen ein Indikator für die Fruchtbarkeit des Explikats. Einige solcher Aussagen ließen sich womöglich aus der Definition instrumentellen Verstehens ableiten – etwa über den Einfluss verschiedener Kontextfaktoren auf die Bedingungen instrumentellen Verstehens. Für den vorliegenden Gegenstandsbereich scheinen hingegen partikuläre Aussagen relevanter zu sein – nämlich darüber, ob ein Stakeholder in einem gegebenen Kontext instrumentell versteht, also einen epistemischen Zustand besitzt, der ihm

die Erfüllung seines Desideratums ermöglicht. Wie oben geschildert, bin ich zuversichtlich, dass das von mir vorgeschlagene Explikat dieses Kriterium erfüllt.

Empirische und informatische XAI-Forschung Das nächste Fruchtbarkeitskriterium besteht darin, informativ für empirische und informatische Forschung zu Erklärbarkeit zu sein. Ein philosophischer Beitrag ist in dieser Hinsicht naturgemäß eingeschränkt, doch weist mein Vorschlag einige interessante Implikationen für benachbarte Disziplinen auf: Instrumentelles Verstehen erfordert ein Überzeugungssystem über systembezogene Aspekte. Während Erklärbarkeit im engen Sinne vornehmlich darauf abzielt, die lokalen oder globalen Entscheidungskriterien verschiedener KI-Systeme aufzudecken, legt meine Definition nahe, dass diese Auffassung zu restriktiv ist. Die systembezogenen Aspekte, die Aufschluss darüber geben, ob ein Desideratum erfüllt ist, sind weitaus vielfältiger (siehe Kapitel 3 und Abschnitt 6.3). Beispielsweise könnten dazu auch Informationen über die verwendeten Trainingsdaten, den Anwendungskontext, den Entwicklungsprozess und die währenddessen getroffenen Entscheidungen zählen. Das passt zu jüngeren Debattenbeiträgen, die Erklärbarkeit auf verschiedene Phasen des Lebenszyklus eines KI-Systems ausdehnen (etwa Deck et al., 2024; Ratti & Graves, 2022). Informationen zu diesen erweiterten systembezogenen Aspekten sind potenziell hilfreiche Inhalte von Verstehensermöglichern.

Eine weitere Implikation ergibt sich aus der Anforderung, dass das Überzeugungssystem eines instrumentell verstehenden Stakeholders gerechtfertigt und moderat faktiv sein muss. Um der Erfüllung dieser Bedingungen zuträglich zu sein, müssen Verstehensermöglicher zuverlässig Auskunft über systembezogene Aspekte geben. XAI-Methoden, die völlig unabhängig vom zu erklärenden System sind und ein hohes Risiko bergen, irreführende Erklärungen zu produzieren (siehe Abschnitt 3.2.1 und 10.1), scheiden daher als geeignete Verstehensermöglicher aus. Weitere wünschenswerte Eigenschaften von Verstehensermöglichern diskutiere ich in Kapitel 12. In Kombination mit meinem Begriff instrumentellen Verstehens könnten diese Eigenschaften herangezogen werden, um bestehende XAI-Methoden aus erkenntnistheoretischer Perspektive zu evaluieren und womöglich auch die Erklärungen neuer Ansätze epistemisch vorteilhafter zu gestalten.

Schließlich bietet mein Vorschlag einen interessanten Ansatzpunkt für die empirische Evaluation der Effektivität von Erklärbarkeitsmethoden und anderen Verstehensermöglichern. Die öffentliche Variante vieler verstehensbezogener Fähigkeiten scheint für die Operationalisierung von instrumentellem Verstehen geeignet. Tatsächlich wurden Fähigkeiten der meisten Fähigkeiten-Cluster bereits genutzt, um XAI-Methoden zu evaluieren (Speith et al., 2024, S. 8–9). Die zusätzlichen, in Abschnitt 9.2 zusammengestellten Fähigkeiten bieten ein vielfältiges Repertoire, das zur Überprüfung des kontextuell relevanten Verstehens genutzt werden könnte.

Großer Geltungsbereich Ein weiteres Kriterium ist ein möglichst großer Geltungsbereich des Explikats. In gewisser Hinsicht ist der Geltungsbereich meiner Definition weitaus

enger als jener des Explikandums, insofern sie nicht alle Fälle von Verstehen *simpliciter* abdeckt. Die Bedingungen instrumentellen Verstehens taugen z. B. nicht dazu, Verstehen als Selbstzweck oder außerhalb von XAI-Kontexten zu analysieren. Allerdings ergibt sich diese Beschränkung unmittelbar aus den theoretischen Kontexten, auf die das Explikat zugeschnitten ist, und den theoretischen Zwecken, die es erfüllen soll.

In anderer Hinsicht besitzt mein Explikat hingegen einen großen Geltungsbereich: Wie im Verlauf dieser Arbeit immer wieder deutlich wurde, sind XAI-Kontexte äußerst vielfältig. Das liegt an den zahllosen Anwendungen, für die KI mittlerweile eingesetzt wird, an den unterschiedlichen, potenziell involvierten Stakeholdern und deren vielfältigen Desiderata sowie an weiteren Kontextfaktoren, die sich auf das Zusammenspiel von Erklärbarkeit, Verstehen und Desiderata auswirken (Langer et al., 2021). Mein Begriff instrumentellen Verstehens soll sich grundsätzlich in all diesen Kontexten als nützlich erweisen, auch wenn es unmöglich ist, diesen großen Geltungsbereich in meiner Diskussion abzudecken. Ein Nachteil dieses Anspruchs ist die wohl unvermeidliche Vagheit meiner Definition.

In der Gesamtschau scheint meine Explikation erfolgreich gewesen zu sein: Ich habe die meist unspezifischen Verweise auf Verstehen als Mediator in XAI-Kontexten in Form einer Definition instrumentellen Verstehens präzisiert. Da diese Definition auf philosophischen Erkenntnissen zu Verstehen *simpliciter* aufbaut, wahrt sie die erforderliche semantische Ähnlichkeit zum Explikandum. Instrumentelles Verstehen ermöglicht die epistemische und mitunter auch substanzielle Erfüllung von gesellschaftlichen Desiderata. Weiterhin erklärt mein Begriffsvorschlag diese Mediatorrolle und wird somit den relevanten theoretischen Zwecken gerecht. Darüber hinaus bietet mein Explikat verschiedene Anknüpfungspunkte für philosophische, informatische und empirische Forschung zu Verstehen in XAI-Kontexten. Das wird im folgenden Kapitel erneut relevant: Ich arbeite nun wünschenswerte Eigenschaften von Verstehensermöglichern heraus, die dem Erwerb instrumentellen Verstehens in XAI-Kontexten zuträglich sind.

Teil III

Verstehensermöglicher

12 Instrumentelles Verstehen und Exemplifikation

Durch meinen Begriff instrumentellen Verstehens habe ich den epistemischen Zustand definiert, der in XAI-Kontexten die epistemische und mitunter auch substanzielle Erfüllung von Desiderata ermöglicht. Wir können uns nun der Frage danach zuwenden, wie dieses Verstehen erworben werden kann. In Kapitel 3 habe ich vorgeschlagen, nicht bloß die Ausgaben von XAI-Ansätzen oder Erklärungen im philosophischen Sinne als Vehikel für den Verstehenserwerb in Betracht zu ziehen. Darüber hinaus kommen auch diverse andere Repräsentationen in Frage (z. B. die Erklärung einer Expertin, die technische Dokumentation eines KI-Systems oder visuelle Repräsentationen). Ich fasse diese Vehikel unter dem Schlagwort ‚Verstehensermöglicher‘ zusammen. Dieses Kapitel widmet sich Verstehensermöglichere und soll dadurch die zweite Forschungsfrage beantworten, die ich in Abschnitt 1.1 formuliert habe:

- (b) Welche Eigenschaften muss ein Verstehensermöglicher aufweisen, um der Erlangung instrumentellen Verstehens zuträglich zu sein?

Eine auf den ersten Blick plausible Antwort lautet, dass in dieser Hinsicht vorteilhafte Verstehensermöglicher möglichst *akkurat* sind: Um (instrumentelles) Verstehen hervorzubringen, sollten Verstehensermöglicher demnach systembezogene Aspekte so wahrheitsgetreu wie möglich repräsentieren. Sollte dies stimmen, wären viele der derzeit verwendeten Post-hoc-XAI-Methoden *aus Prinzip* defizitär. Dies lässt sich anhand von LIME verdeutlichen (siehe Abschnitte 3.2.1 und 7.2): LIME berechnet eine lineare Funktion, welche die (potenziell nicht-lineare) Funktion des zugrunde liegenden KI-Modells lokal approximiert. Diese Approximation wird dann genutzt, um eine lokale Erklärung für eine bestimmte Ausgabe des KI-Systems zu generieren (Ribeiro et al., 2016). Allein aufgrund dieses Funktionsprinzips unterscheidet sich die erklärende zwangsläufig von der erklärten Funktion. Die resultierende Erklärung ist daher in manchen Hinsichten inakkurat. Diesen Punkt formuliert Cynthia Rudin in einem einflussreichen Aufsatz:

[Post-hoc] [e]xplanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable.) (Rudin, 2019, S. 207)

Unter anderem auf Basis dieser Kritik plädiert Rudin für ante-hoc erklärbare Systeme und gegen die Verwendung von opaken Systemen, die lediglich durch Post-hoc-Ansätze transparent gemacht werden können – wenigstens in Kontexten, in denen viel auf dem Spiel steht (Rudin, 2019). In der Tat steht zu befürchten, dass manche XAI-Methoden irreführende Erklärungen generieren, die wenig mit dem Entscheidungsprozess des erklärten KI-Systems zu tun haben (siehe Abschnitt 3.2.1). Doch scheint Rudin einen grundlegenden Punkt zu formulieren: Eine Abweichung zwischen Erklärung und Modell sei *per se* problematisch, daher sollten wir in XAI-Kontexten danach streben, vollständig akkurate Erklärungen zu produzieren. Anders ausgedrückt: Wir sollten KI-Systeme als ihre eigene Erklärung verwenden. Im Kontext dieser Arbeit lässt sich Rudins übergeordneter Punkt aus mindestens zwei zusammenhängenden Gründen in Zweifel ziehen.

Erstens ist zumindest der hier betrachtete epistemische Zustand – also instrumentelles Verstehen – mit falschen Überzeugungen und Verstehenslücken kompatibel. Instrumentelles Verstehen entspricht keineswegs maximalem Verstehen. Wenn inakkurate Erklärungen zu fehlenden, falschen oder nur approximativ wahren Überzeugungen führen, mindert dies zwar den Verstehensgrad *simpliciter*, beeinträchtigt aber nicht zwangsläufig das instrumentelle Verstehen des fraglichen Stakeholders: Sofern die kontextuell zentralen Überzeugungen zumindest approximativ wahr sind, ist die Faktizitätsbedingung instrumentellen Verstehens erfüllt. Abweichungen zwischen Modell und Erklärung gefährden daher nicht automatisch instrumentelles Verstehen.

Zweitens wurde in Abschnitt 8.3.1 deutlich, dass auch inakkurate Verstehensermöglicher streng faktives Verstehen hervorbringen können. Kompetente Wissenschaftler:innen erlangen mithilfe des Idealen-Gas-Modells keine falschen Überzeugungen, sondern interpretieren das Modell derart, dass sie lediglich wahre Überzeugungen formen. Analog kann auch eine partiell falsche oder nur approximativ wahre XAI-Erklärung korrekt interpretiert werden und den Ausgangspunkt für ausschließlich wahre Überzeugungen über systembezogene Aspekte bilden. Daher führen inakkurate Erklärungen nicht *per se* zu falschen Überzeugungen und beeinträchtigen nicht notwendigerweise das resultierende Verstehen.

Im Gegenteil: Gezielte Idealisierungen oder Abstraktionen können dem Erwerb wahrer Überzeugungen sogar zuträglich sein. Eine besonders schöne Variante dieses Gedankens formuliert Lewis Carroll in seinem Roman *Sylvie and Bruno Concluded*. Darin unterhält sich der Ich-Erzähler mit ‚Mein Herr‘, einem Reisenden aus einem bizarren Land:

“What a useful thing a pocket-map is!” I remarked.

“That’s another thing we’ve learned from *your* Nation,” said Mein Herr, “map-making. But we’ve carried it much further than *you*. What do you consider the largest map that would be really useful?”

“About six inches to the mile.”

“Only *six inches*!” exclaimed Mein Herr. “We very soon got to six *yards* to the mile. Then we tried a *hundred* yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of *a mile to the mile*!”

“Have you used it much?” I enquired.

“It has never been spread out, yet,” said Mein Herr: “the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well. [...]” (Carroll, 1894, S. 169; Hervorhebung im Original)

Diese Passage führt Rudins Argument *ad absurdum*: Offensichtlich sind Abweichungen zwischen epistemischen Repräsentationen und der Wahrheit mitunter wünschenswert. Die Gründe dafür sind vornehmlich pragmatisch, aber epistemisch relevant: Wir wollen viele äußerst komplexe Dinge verstehen – dazu zählen nicht nur neuronale Netze, sondern auch viele andere Verstehensobjekte (zum Beispiel unsere natürliche Umwelt). Gleichzeitig sind unsere zeitlichen und kognitiven Ressourcen begrenzt. Zielgerichtete Vereinfachungen sind mitunter der einzige Weg, um solche komplexen Dinge zumindest ein Stück weit verständlich zu machen. Anstatt Abstraktionen oder Idealisierungen allein aufgrund der daraus resultierenden Abweichung von der Wirklichkeit abzulehnen – und damit das Verstehen tiefer neuronaler Netze und anderer epistemisch opaker Systeme weitestgehend aufzugeben –, können wir Vereinfachungen auch als Werkzeug begreifen, mit dem sich die architektonische Opazität moderner KI-Systeme handhabbar machen lässt (vgl. Abschnitt 2.3.1; diese Position teilen Fleisher, 2022 und Páez, 2019).

Überlegungen wie diese werden in der Wissenschafts- und Erkenntnistheorie ausführlich diskutiert. Entsprechend heben auch viele andere Autor:innen hervor, dass Idealisierungen, Abstraktionen, Fiktionen, Approximationen oder andere Falschheiten dem Verstehenserverwerb mithilfe z. B. wissenschaftlicher Repräsentationen zuträglich sein können (siehe etwa Bokulich, 2016; Cooper, 1995; de Regt, 2017; de Regt & Dieks, 2005; de Regt & Gijsbers, 2017; Elgin, 2017, 2022; Lawler, 2021; Potochnik, 2015; Rohwer & Rice, 2013). Catherine Elgin hat dafür den Begriff der *gelungenen Falschheiten* (siehe Fußnote 93) geprägt:

[...][Felicitous] falsehoods are inaccurate in ways that enable them to non-accidentally provide epistemic access to obscure or occluded aspects of their targets. Not despite, but because of their inaccuracy, they afford the access that they do. When inaccurate models facilitate reasoning about the phenomena, they are epistemically felicitous. (Elgin, 2022, S. 12)

Übertragen auf meine Forschungsfrage bedeutet dies, dass vorteilhafte Verstehensermöglicher gelungene Falschheiten beinhalten. Solche gelungenen Falschheiten verschaffen einem Stakeholder in XAI-Kontexten epistemischen Zugang zu kontextuell relevanten systembezogenen Aspekten und unterstützen ihn dabei, wahre oder annähernd wahre Überzeugungen zu erwerben, die zentral für die Erfüllung seines Desideratums sind. Doch was macht die Falschheiten eines Verstehensermöglichers zu *gelungenen Falschheiten*?

Elgins Vorschlag lautet, dass sich gelungene Falschheiten dadurch auszeichnen, dass sie relevante Eigenschaften *exemplifizieren* (Elgin, 2017, S. 5). In diesem Kapitel möchte ich mithilfe von Elgins Theorie der Exemplifikation einige Eigenschaften von Verstehensermöglichern herausarbeiten, die den Erwerb instrumentellen Verstehens befördern. Ich

nenne solche Verstehensermöglicher ‚EXEMPLARE‘ und grenze sie von epistemisch weniger vorteilhaften ‚SAMPLES‘ ab. Meinem Vorschlag zufolge sind EXEMPLARE, nicht aber SAMPLES, effektive Verstehensermöglicher für den Erwerb instrumentellen Verstehens in XAI-Kontexten. Diesen Vorschlag habe ich in ähnlicher Form bereits im Rahmen eines publizierten Aufsatzes dargelegt (Mann, 2025), entsprechend gibt das verbleibende Kapitel in weiten Teilen den Inhalt dieses Aufsatzes wieder.¹³⁰

Vorab eine terminologische Anmerkung: Obwohl ‚Beispiel‘, ‚Sample‘ und ‚Exemplar‘ (insbesondere im Englischen) oft synonym verwendet werden, werde ich die in Kapitälchen gedruckten Ausdrücke ‚EXEMPLAR‘ und ‚SAMPLE‘ fortan ausschließlich im soeben beschriebenen Sinne gebrauchen. Der Ausdruck ‚Beispiel‘ bezieht sich hingegen auf jede Instanz, die einige ihrer Merkmale exemplifiziert; dies umfasst sowohl EXEMPLARE als auch SAMPLES. Damit weiche ich von Elgins eigener Begriffsverwendung ab, derzufolge ‚*exemplar*‘ als Oberbegriff dient (Elgin, 2017, S. 184).

Ich illustriere meinen Vorschlag nachfolgend anhand von Verstehensermöglichere aus dem Bereich Bildklassifikation: visuelle Erklärungen der Klassifikation eines oder mehrerer Bilder als Mitglied der Bildklasse ‚Rotkehlchen‘.¹³¹ Ich gehe wie folgt vor: In Abschnitt 12.1 lege ich Elgins Theorie der Exemplifikation in ihren Grundzügen dar. In Abschnitt 12.2 beschreibe ich fünf wünschenswerte Eigenschaften von EXEMPLAREN und grenze diese von den epistemisch unvorteilhaften SAMPLES ab. Außerdem erläutere ich, wie EXEMPLARE zur Erfüllung der Bedingungen instrumentellen Verstehens beitragen. In Abschnitt 12.3 übertrage ich diese Überlegungen auf verschiedene XAI-Methoden und prüfe, ob deren Erklärungen in Hinblick auf ein mögliches epistemisches Ziel in XAI-Kontexten als EXEMPLARE gelten können. Schließlich antworte ich in Abschnitt 12.4 auf drei Einwände.

12.1 Zu Elgins Theorie der Exemplifikation

Elgin zufolge ist die Exemplifikation einer der primären Modi der Referenz (neben der Denotation, siehe unten). Ein Beispiel fungiert als Symbol,¹³² das auf einige der Eigenschaften, Muster oder Relationen verweist bzw. referenziert, die es instanziiert (ebd.). Aufgrund

¹³⁰Ich möchte nochmals Claus Beisbart und einem anonymen Reviewer bei *AI & Society* für äußerst hilfreiche Anmerkungen zu früheren Versionen des Aufsatzes danken. Abgesehen von meinem eigenen Aufsatz hat Elgins Theorie der Exemplifikation in der (philosophischen) XAI-Forschung bislang fast keine Aufmerksamkeit erfahren. Eine Ausnahme ist ein Beitrag von Andrés Páez (2019, S. 448–449), der kurz auf den Mehrwert von Exemplifikation für Verstehen in XAI-Kontexten eingeht. Elgins eigene Forschung befasst sich meines Wissens nicht mit KI oder Erklärbarkeit.

¹³¹Mit ‚Bildklasse‘ meine ich eine Kategorie von Bildern (z. B. Rotkehlchenfotos), keine Objektklasse (z. B. Rotkehlchen). Darüber hinaus zielt eine XAI-Erklärung darauf ab, die wahren Ursachen oder Gründe für die Ausgabe eines KI-Systems zu erhellen. Weder muss die zugewiesene Bildklasse mit der korrekten Kategorisierung dieses Bildes übereinstimmen, noch müssen die Gründe für die Klassifikation tatsächlich für diese Klassifikation sprechen.

¹³²Elgin hat einen breiten Symbolbegriff, der sowohl sprachliche als auch nichtsprachliche Dinge umfasst – einschließlich Kunstwerke oder wissenschaftliche Modelle.



Abbildung 12.1: Beispiel der Bildklasse ‚Rotkehlchen‘. Wir können für unsere Zwecke stipulieren, dass die durch das Beispiel exemplifizierten Merkmale jene Pixelgruppen sind, die einem roten Brust- und Kopfgefieder entsprechen. Es sind diese Merkmale, die auf den Referenten (z. B. typische Mitglieder der Bildklasse ‚Rotkehlchen‘) imputiert werden. Die Extension jener Instanzen, die diese Merkmale teilen, wird dann durch das Beispiel typisiert. Bildquelle: Public Domain. Siehe Teil IV für Details.

dieser Referenz ist ein Beispiel immer ein Beispiel für etwas. Beispielsweise kann ein Bild wie Abbildung 12.1 ein Beispiel für die Bildklasse ‚Rotkehlchen‘ sein. Gleichmaßen könnte dasselbe Bild ein Beispiel für ein bestimmtes Merkmal von Rotkehlchen sein, z. B. für den spezifischen Rotton der Brust und des Kopfes (Mann, 2025, S. 39). Wenn ein Beispiel einige seiner Merkmale exemplifiziert, macht es diese Merkmale deutlich, hebt sie hervor bzw. zeigt sie auf (Elgin, 2017, S. 185). Wie Elgin es ausdrückt, *imputiert* das Beispiel die exemplifizierten Merkmale auf dasjenige, wofür es ein Beispiel ist (ebd., S. 266). Dies erlaubt der Interpretin des Beispiels, diese Merkmale auch seinem Referenten zuzuschreiben. Das Beispiel bietet dadurch epistemischen Zugang zu bestimmten Merkmalen seines Referenten (vgl. ebd., S. 253). Exemplifikation hebt dabei nicht nur ohnehin saliente Merkmale hervor, sondern kann auch eine Quelle für Salienz sein (ebd., S. 187).

Im Folgenden werde ich Elgins Theorie der Exemplifikation näher charakterisieren. Eine erschöpfende Darstellung würde den vorliegenden Rahmen sprengen,¹³³ daher konzentriere ich mich auf fünf Aspekte, die ich für die Exemplifikation in XAI-Kontexten als relevant erachte: Instanziierung, Referenz, Typisierung, Selektivität und Interpretation.

¹³³Elgin publiziert bereits seit Jahrzehnten zu diesem Thema, und im Laufe der Zeit hat sich ihre Theorie teils deutlich verändert. Ich stütze mich nachfolgend auf Elgins Darstellung der Exemplifikation in *True Enough* (2017).

Erstens: Exemplifikation erfordert Instanziierung. Um charakteristische Merkmale von Mitgliedern der Bildklasse ‚Rotkehlchen‘ zu exemplifizieren, muss ein Foto diese Merkmale instanzieren – zum Beispiel in Form von Pixeln, die einem rötlichen Brust- und Kopfgefieder entsprechen (Mann, 2025, S. 39). In diesem Punkt unterscheidet sich die Exemplifikation von der Denotation: Denotation ist stipulativ (Elgin, 2017, S. 253–255). Die Aussage ‚Dieser Bleistift repräsentiert ein Rotkehlchen‘ genügt, damit der Bleistift den Vogel denotiert, aber sie reicht nicht aus, um ihn zu exemplifizieren – schließlich ist ein Bleistift keine Instanz eines Rotkehlchens (vgl. ebd., S. 185). Der Bleistift kann jedoch bestimmte Merkmale von Rotkehlchen instanzieren (etwa die besondere Farbe des Brustgefieders) und sie dadurch auch exemplifizieren. Ein Beispiel, das Merkmale seines Referenten exemplifiziert, *instanziiert* also diese Merkmale (Mann, 2025, S. 39).

Zweitens: Bloße Instanziierung ist nicht hinreichend für Exemplifikation. Es gibt unzählige Bilder von Rotkehlchen. Wenn sie nicht auf die Bildklasse ‚Rotkehlchen‘ referenzieren, sind sie keine Beispiele dafür (siehe Elgin, 2017, S. 185). Nehmen wir ein Bild, das ein Rotkehlchen auf einer Birke zeigt. In einem Kontext kann sich dieses Bild auf Rotkehlchen beziehen – etwa, wenn eine Ornithologin das Foto verwendet. In einem anderen Kontext könnte dasselbe Bild hingegen ausschließlich auf Birken verweisen – zum Beispiel, wenn eine Försterin anhand des Bildes die Form von Birkenblättern erläutert. Im letzteren Fall ist das Bild kein Beispiel für Rotkehlchen, weil es nicht darauf verweist. Ein solcher Verweis ist nicht schwer zu erzielen. Es reicht z. B. aus, wenn man das Foto mit der Beschreibung ‚Rotkehlchen‘ versieht (vgl. ebd., S. 255). Bei XAI-Erklärungen geht normalerweise aus dem Kontext hervor, worauf die fragliche Erklärung verweist (etwa auf die ausschlaggebenden Merkmale für eine bestimmte Klassifikation). Zusammengefasst: Ein Beispiel, das Merkmale seines Referenten exemplifiziert, *referenziert* bzw. *verweist* auf diese Merkmale (Mann, 2025, S. 39).

Drittens: Ein einzelnes Beispiel ist deswegen aufschlussreich, weil es über sich selbst hinausweist. Wenn ein Rotkehlchenfoto Pixelgruppen exemplifiziert, die einem rötlichen Brustgefieder entsprechen, dann fungiert es als Beispiel, das auf die Extension genau jener Bilder verweist, die diese Merkmale teilen. In Elgins Worten: Das Beispiel typisiert diese Extension (ebd.). Durch die Typisierung macht das Beispiel deutlich, in welcher Hinsicht sich die Instanzen dieser Extension ähneln (Elgin, 2017, S. 263–267). Ein Beispiel, das Merkmale seines Referenten exemplifiziert, *typisiert* also die Extension aller und ausschließlich jener Instanzen, die diese Merkmale teilen (Mann, 2025, S. 39).

Viertens ist Exemplifikation selektiv. Das Exemplifizieren bestimmter Merkmale erfordert das Hervorheben dieser Merkmale. Ein Beispiel kann zwar mehrere der von ihm instanziierten Eigenschaften hervorheben, aber nicht alle auf einmal (Elgin, 2017, S. 185). Stattdessen führt die Hervorhebung oder Betonung bestimmter Merkmale dazu, dass andere weggelassen, heruntergespielt oder anderweitig in den Hintergrund gerückt werden. Die Illustration eines Rotkehlchens in einem Bestimmungsbuch kann z. B. eine rötliche Brust betonen, während eine bestimmte Körperhaltung oder -größe zwar instanziiert, aber nicht exemplifiziert wird (Mann, 2025, S. 39). Selektivität ist nicht unbedingt problematisch, da

ein Beispiel nur relevante Merkmale exemplifizieren muss, um das gewünschte Verständnis hervorzubringen. Welche Merkmale als relevant gelten, hängt vom kontextuellen Zweck des Beispiels ab (Elgin, 2017, S. 193–194). Zusammengefasst: Ein Beispiel, das Merkmale seines Referenten exemplifiziert, hebt diese Merkmale *selektiv* hervor, während andere Eigenschaften in den Hintergrund gerückt oder sogar gänzlich weggelassen werden (Mann, 2025, S. 39).

Fünftens muss ein Beispiel in mindestens zwei Hinsichten richtig interpretiert werden, um Verstehen hervorzubringen. Beide haben mit der Bestimmung der Extension zu tun, die ein Beispiel typisiert (ebd., S. 40). Betrachten wir diese Formen der Interpretation etwas genauer.

Zunächst einmal muss die Interpretin erkennen, welche Eigenschaften das Beispiel exemplifiziert – also welche Merkmale auf den Referenten imputiert werden bzw. wodurch sich die typisierte Extension auszeichnet (Elgin, 2017, S. 188). Denn nicht alle Eigenschaften, die ein Beispiel instanziiert, werden exemplifiziert. Elgin unterscheidet zwischen inerten, stützenden (engl. *scaffolding*) und exemplifizierten Merkmalen (ebd., S. 265–266). *Inerte Merkmale* sind einfach irrelevant, z. B. die Schriftart, in der das Wort ‚Rotkehlchen‘ abgedruckt ist. *Stützende Merkmale* werden nicht selbst exemplifiziert, sondern ermöglichen die Exemplifikation. Beispielsweise kann die Abbildung in einem Bestimmungsbuch kleine Pfeile enthalten, die auf charakteristische Merkmale einer Vogelart hinweisen (wie in Peterson, 1980). Schließlich sind *exemplifizierte Merkmale* jene Eigenschaften, die auf den Referenten imputiert werden. Wird das Beispiel korrekt interpretiert, gibt es der Interpretin Aufschluss über diese Aspekte des Referenten.

Zur Veranschaulichung: Ein durchschnittliches Beispiel der Bildklasse ‚Rotkehlchen‘ instanziiert zahllose Merkmale, die für Mitglieder dieser Klasse nicht charakteristisch sind. Angenommen, die korrekte Interpretation dieses Beispiels besteht darin, als exemplifiziertes Merkmal ein rötliches Brustgefieder (bzw. korrespondierende Pixelgruppen) zu identifizieren. Diese Interpretation ermöglicht die Bestimmung der Extension, die das Beispiel typisiert – nämlich alle Bilder, die dieses Merkmal aufweisen (Mann, 2025, S. 40). Wird den anderen Mitgliedern der Bildklasse hingegen die Farbe des Hintergrunds oder der Schwanzfedern zugeschrieben, schlägt die Interpretation fehl: Der Interpretin des Beispiels ist es nicht gelungen, exemplifizierte von inerten Merkmalen zu unterscheiden.

Die Identifikation exemplifizierter Merkmale kann durch diverse Interpretationshilfen unterstützt werden, welche die Mehrdeutigkeit des Beispiels reduzieren (Elgin bezeichnet dies als ‚*stage setting*‘; Elgin, 2017, S. 192). Zum Beispiel kann eine Ornithologin auf charakteristische Merkmale eines Rotkehlchens hinweisen, wenn sie ihren Zuhörer:innen einen solchen Vogel präsentiert (vgl. ebd., S. 188–189). Relevante Kontraste können die Interpretation ebenfalls erleichtern (ebd., S. 188–192). Wenn man beispielsweise ein Rotkehlchen mit einem Zwergschnäpper kontrastiert, kann dies helfen, die jeweils distinkten Merkmale der beiden Arten abzuleiten (Mann, 2025, S. 40). Auch die Verwendung eines Beispiels, das von sich aus leichter zu interpretieren ist, stellt eine wichtige Interpretationshilfe dar. Man kann ein typisches Beispiel wählen, bei dem die relevanten Eigenschaften deutlich

ausgeprägt sind und hervorstechen (Elgin, 2017, S. 192). Es kann sogar sinnvoll sein, ein neues Beispiel zu kreieren, um bestimmte Merkmale von vorneherein hervorzuheben. Elgins Standardbeispiel ist der Farbfächer eines Farbenherstellers, der bei der Auswahl einer Wandfarbe verwendet werden kann (ebd., S. 187–188). Der Farbfächer instanziiert nur die Merkmale der Farbe, die exemplifiziert werden sollen (d. h. den Farbton) – nicht aber kontextuell irrelevante Eigenschaften wie ihre Konsistenz oder ihren Geruch. Zusammen mit dem üblichen Hintergrundwissen über solche Farbfächer trägt der selektive Charakter des Beispiels maßgeblich zu seiner korrekten Interpretation bei (Mann, 2025, S. 40).

Hier kommt eine Uneindeutigkeit in Elgins Darstellung zum Tragen. Einerseits werden exemplifizierte Merkmale dadurch charakterisiert, dass sie durch das Beispiel selektiv hervorgehoben werden und andere Merkmale herunterspielen (z. B. Elgin, 2017, S. 185, 188). Demzufolge würde ein Merkmal nur dann exemplifiziert, wenn es hervorgehoben wird. Exemplifizierte Eigenschaften wären dann jene, die durch das Beispiel hervorgehoben werden. Die korrekte Interpretation wäre dadurch so gut wie garantiert. Andererseits diskutiert Elgin Fälle, in denen Interpret:innen daran scheitern, exemplifizierte Merkmale zu identifizieren (etwa ebd., S. 187–189, 253), und unterscheidet effektive von ineffektiven Beispielen (z. B. ebd., S. 192). Demzufolge würden exemplifizierte Merkmale nicht zwangsläufig hervorgehoben und könnten daher schwierig von inerten oder stützenden Merkmalen zu unterscheiden sein. Dass ein Beispiel ein Merkmal exemplifiziert, würde nicht sicherstellen, dass es korrekt interpretiert wird. Der ersten Lesart zufolge ist Salienz also notwendig für Exemplifikation, der zweiten Auslegung zufolge nicht.

Soweit ich sehen kann, lässt sich diese Uneindeutigkeit nur auflösen, indem man einen Teil von Elgins Aussagen relativiert. Die folgende Variante der zweiten Lesart scheint die geringste Abweichung gegenüber Elgins Darstellung zu erfordern: Demnach ‚soll‘ ein Beispiel im fraglichen Kontext einige seiner Eigenschaften salient machen (etwa, weil die beispielgebende Person diese Absicht verfolgt, oder weil das Beispiel auf eine gewisse Weise erzeugt wurde). Bei diesen Eigenschaften handelt es sich um die exemplifizierten Merkmale. Salienz ist aber auch abhängig von der Interpretin des Beispiels. Wenn eine Ornithologin einer Schulkasse ein Rotkehlchen zeigt, gibt es eine korrekte Interpretation dieses Beispiels, weil der Vogel in diesem Kontext spezifische Eigenschaften von Rotkehlchen exemplifiziert. Ob es den Schüler:innen gelingt, diese Eigenschaften zu identifizieren, steht aber auf einem anderen Blatt (ebd., S. 188–189). Entscheidend ist: Auch wenn die Interpretation fehlschlägt und die relevanten Eigenschaften für die Schulkinder nicht salient werden, exemplifiziert das Beispiel bestimmte Merkmale von Rotkehlchen.

Ich stütze mich nachfolgend auf diese Lesart. Für mich hängt aber nicht viel daran, welcher Auslegung man folgt. Wäre Salienz notwendig für Exemplifikation, würden die folgenden Überlegungen zur Interpretation von Beispielen und verschiedenen Interpretationshilfen eine Vorbedingung von Exemplifikation betreffen. Bei der ersten Form der Interpretation ginge es dann nicht um die Identifikation exemplifizierter Merkmale, sondern um die Bestimmung von relevanten Merkmalen, die das Beispiel mit seinem Referenten teilt. Erst wenn diese (für eine Interpretin) salient sind, würden sie exemplifiziert.

Kommen wir zur zweiten Form der Interpretation von Beispielen. Die Interpretin muss auch erkennen, was das für eine Extension ist, die das Beispiel typisiert. Ein Beispiel typisiert die Extension der Instanzen, welche die von ihm exemplifizierten Merkmale teilen – so weit, so gut. Aber welche Extension ist das (ebd., S. 190)? Diese Form der Interpretation hat mit der Konkretisierung des Referenten zu tun: Nehmen wir an, ein Rotkehlchenbild typisiert die Extension von Bildern mit Pixeln, die einem rötlichen Brust- und Kopfgefieder entsprechen. Um diese Information nutzen zu können, muss man wissen, ob es sich dabei um die Extension der typischen oder gar aller Rotkehlchenbilder handelt, oder ob vielleicht nur eine Minderheit der Rotkehlchenbilder dieses Merkmal teilt (Mann, 2025, S. 40). Diese Information wird meist nicht durch das Beispiel selbst bereit gestellt. Welche Extension es repräsentiert, hängt davon ab, wie das Beispiel ausgewählt oder generiert wurde (Elgin, 2017, S. 190) – in XAI-Kontexten also z. B. von der verwendeten Erklärbarkeitsmethode.

Zusammengefasst: Damit von einem Beispiel auf die Extension der Instanzen geschlossen werden kann, die es typisiert, muss es in mindestens zwei verschiedenen Hinsichten korrekt *interpretiert* werden (Mann, 2025, S. 40).

Instanziierung, Referenz, Typisierung, Selektivität und Interpretationsbedarf zeichnen alle Beispiele aus. Allerdings können Beispiele daran scheitern, der Adressatin den gewünschten epistemischen Zugang zu exemplifizierten Merkmalen zu verschaffen: Erstens können Beispiele irreführend sein. Sie können vorgeben, Merkmale des Referenten zu exemplifizieren, die er in Wirklichkeit nicht besitzt (Elgin, 2017, S. 199). Das Bild eines Rotkehlchens, das Rotkehlchen ein blaues Brustgefieder zuschreibt, imputiert ein Merkmal, das der Referent nicht aufweist (Mann, 2025, S. 40). Zweitens können Beispiele falsch interpretiert werden (Elgin, 2017, S. 189), etwa, wenn die Interpretin Merkmale des Beispiels auf den Referenten imputiert, die der Referent nicht besitzt. Die Leserin eines Bestimmungsbuchs, die annimmt, dass eine Rotkehlchenabbildung die Zweidimensionalität von Rotkehlchen exemplifiziert, interpretiert das Beispiel falsch (Mann, 2025, S. 40). Drittens kann die Interpretin den Referenten des Beispiels fehldeuten (vgl. Elgin, 2017, S. 190). Jemand könnte beispielsweise glauben, dass ein bestimmter Vogel ein Rotkehlchen ist, obwohl es sich in Wirklichkeit um einen Gartenrotschwanz handelt. Kurz: Nicht alle Beispiele sind gute Beispiele, und nicht alle Interpret:innen sind kompetent (Mann, 2025, S. 40).

12.2 SAMPLES, EXEMPLARE und instrumentelles Verstehen

Soweit zu Elgins Theorie der Exemplifikation. Doch was hat es mit meiner Unterscheidung zwischen EXEMPLAREN und SAMPLES auf sich? Der Unterschied lässt sich anhand des folgenden Paares an Fallbeispielen verdeutlichen:

Schnappschuss: Moira möchte verstehen, woran Ornitholog:innen Rotkehlchen erkennen; sie hat aber keinerlei Vorwissen über das Aussehen dieser Vogelart. Moira sucht online nach Bildern von Rotkehlchen und betrachtet ein zufällig ausgewähltes

Foto. Sie hat Pech – auch wenn Moira nichts davon ahnt, handelt es sich bei dem Vogel auf dem Foto um ein atypisches Rotkehlchen, das eine ungewöhnliche Färbung des Gefieders aufweist. Zudem ist der Vogel von mehreren anderen Vögeln umgeben, die Moira ebenfalls nicht kennt.

Fachbuch: Auch Arda hat keine Ahnung, wie ein Rotkehlchen aussieht. Im Gegensatz zu Moira konsultiert Arda ein Vogelbestimmungsbuch, das eine didaktisch aufbereitete Abbildung eines Rotkehlchens enthält. Diese Abbildung zeigt nicht nur ein typisches Rotkehlchen. Zusätzlich werden dessen wichtigste Merkmale hervorgehoben und unspezifische Eigenschaften außen vor gelassen.

Es liegt auf der Hand, dass die Abbildung aus dem *Fachbuch*-Fall hilfreicher ist als das Foto aus *Schnappschuss*, um zu verstehen, was Rotkehlchen auszeichnet. Diese Abbildung *instanziiert* typische Merkmale von Rotkehlchen; durch den Kontext (ein Vogelbestimmungsbuch samt Erläuterungen) *referenziert* sie auf diese Vogelart; die Abbildung hebt *selektiv* relevante Eigenschaften von Rotkehlchen hervor und *typisiert* dadurch jene Extension von Vögeln, die diese Merkmale teilen. Weiterhin ist die Illustration relativ leicht zu *interpretieren*: Dank verschiedener Interpretationshilfen (z. B. Auslassungen, Erläuterungen oder Pfeile) kann Arda exemplifizierte Merkmale vergleichsweise einfach identifizieren; und es ist eindeutig, dass die typisierte Extension typische Rotkehlchen umfasst (anstatt z. B. seltene Varianten). Im Gegensatz dazu instanziiert das unübersichtliche Foto, das Moira heranzieht, viele charakteristische Merkmale dieser Vogelart nicht (z. B. die übliche Färbung des Gefieders). Daher typisiert es auch nicht die Extension typischer Rotkehlchen, für die sich Moira interessiert. Der Schnappschuss hebt nicht selektiv relevante Merkmale hervor, sondern instanziiert eine Vielzahl teils irrelevanter Eigenschaften. Außerdem ist das Foto schwierig zu interpretieren: Welche der instanziierten Merkmale sind relevant, welche kann Moira ignorieren? Welche Extension wird durch das Bild typisiert?

Der Unterschied zwischen dem willkürlich ausgewählten Foto und der Abbildung aus dem Bestimmungsbuch illustriert den Unterschied zwischen EXEMPLAREN und SAMPLES: Das Bild aus *Fachbuch* ist ein EXEMPLAR. Die Illustration exemplifiziert charakteristische Merkmale von Rotkehlchen, ist leicht zu interpretieren und hilft Arda, relevante Eigenschaften dieser Vogelart zu identifizieren. Dadurch verschafft ihm die Abbildung epistemischen Zugang zu diesen Merkmalen. Im Gegensatz dazu ist das unübersichtliche Foto eines untypischen Rotkehlchens aus *Schnappschuss* schwierig zu interpretieren und instanziiert nicht die charakteristischen Merkmale von Rotkehlchen. Es eignet sich nicht gut dafür, Moira epistemischen Zugang zu diesen Merkmalen zu ermöglichen. Kurz: Bei dem Foto handelt es sich um ein SAMPLE (Mann, 2025, S. 38).

Ich habe dieses Beispiel deswegen gewählt, weil im Kontext von Vogelbestimmungsbüchern Überlegungen angestellt wurden, die meinem Vorschlag ähneln. Die Soziologen John Law und Michael Lynch argumentieren im Rahmen ihrer Untersuchung der Praxis der Vogelbeobachtung, dass die schematische Darstellung einer Vogelart besser geeignet ist als ein Foto, um die relevanten Merkmale dieser Art hervorzuheben. Denn eine solche

Abbildung liefere klare Unterscheidungskriterien, indem sie „die meisten möglichen Unterschiede auf kunstvolle Weise als irrelevant“ darstelle (Law & M. Lynch, 1988, S. 284–285; meine Übersetzung).

Die schematischen Darstellungen, auf die sich Law & Lynch beziehen, stammen vom Ornithologen Roger Tory Peterson. Petersons Vogelbestimmungsbücher waren insbesondere deswegen einflussreich, weil er bei den darin enthaltenen Abbildungen neue Wege ging. Unter anderem war er ein Befürworter der Verwendung schematischer Darstellungen zur Vogelbestimmung. Peterson war der Meinung, dass sie besser geeignet seien, um so genannte ‚Marker‘ (engl. *field marks*) hervorzuheben, also die charakteristischen Merkmale einer Art (Mann, 2025, S. 44–45):

A photograph is a record of a fleeting instant; a drawing is a composite of the artist's experience. The artist can edit out, show field marks to best advantage, and delete unnecessary clutter. [...] A photograph is subject to the vagaries of color temperature, make of film, time of day, angle of view, skill of the photographer, and just plain luck. The artist has more options and far more control [...]. [...] Whereas a photograph can have a living immediacy, a good drawing is really more instructive. (Peterson, 1980, S. 9–10)

Wie Schnappschüsse von Vögeln können SAMPLES ‚überflüssiges Klimbim‘ enthalten oder kontextuell relevante Merkmale auslassen. Das macht es für die Interpretin schwierig zu entscheiden, welche Merkmale von Bedeutung sind. Bestenfalls verkomplizieren SAMPLES lediglich den Verstehensserwerb, weil ihre Interpretation aufwändiger ist. Schlimmstenfalls verhindern SAMPLES Verstehen, weil sie falsch interpretiert werden oder kontextuell relevante Merkmale nicht exemplifizieren. Im Gegensatz dazu bezeichne ich Beispiele, die das Risiko der Fehlinterpretation durch geeignete Interpretationshilfen mindern, als EXEMPLARE. EXEMPLARE sind gezielt gestaltet oder ausgewählt, um das kontextuell erforderliche Verstehen hervorzubringen. Sie machen durch Auslassungen, Hervorhebungen oder andere Interpretationshilfen deutlich, welche Merkmale wichtig sind. Ein EXEMPLAR ist ein ‚klarer Fall‘, der den kognitiven Aufwand und das Risiko von Missverständnissen verringert. Mit anderen Worten: EXEMPLARE ähneln den Illustrationen aus Petersons Bestimmungsbüchern (Mann, 2025, S. 45).

Dass die in XAI-Kontexten verwendeten Verstehensermöglicher einfach zu interpretieren sind, ist von besonderer Wichtigkeit. Menschen neigen dazu, Bildklassifikationssysteme (aber auch andere KI-Systeme) zu anthropomorphisieren – sie schreiben ihnen fälschlicherweise ihre eigenen Klassifikationsstrategien zu, obwohl diese Systeme häufig ganz andere Kriterien heranziehen als Menschen (vgl. Mueller, 2020; Pelland et al., 2024; siehe auch Abschnitt 2.3.1). SAMPLES bergen in XAI-Kontexten daher ein besonders großes Risiko für Missverständnisse.

Hier sind einige Klarstellungen (Mann, 2025, S. 45). Erstens: Der wichtigste Unterschied zwischen EXEMPLAREN und SAMPLES besteht in der Schwierigkeit ihrer Interpretation. Oftmals bilden sie keine klar abgegrenzten Kategorien, sondern unterscheiden sich nur

graduell.

Zweitens behaupte ich nicht, dass EXEMPLARE notwendigerweise künstlich erzeugte Instanzen sind. Im Kontext der Bildklassifikation kann z. B. auch ein natürliches Bild ausreichend leicht zu interpretieren sein, um als EXEMPLAR zu gelten. Künstliche Beispiele bieten jedoch mehr Spielraum für die Integration von Interpretationshilfen und sind daher unter Umständen einer ‚natürlichen‘ Instanz vorzuziehen.

Drittens sind EXEMPLARE nicht der richtige Verstehensermöglicher für alle denkbaren epistemischen Zwecke. Es kann z. B. bei der Bildklassifikation durchaus Aspekte geben, die sich nicht einmal durch mehrere EXEMPLARE visualisieren lassen – etwa Klassifikationskriterien für große und heterogene Bildklassen wie ‚Tier‘ oder ‚Obst‘. In anderen Fällen könnten EXEMPLARE nicht für alle potenziellen Adressat:innen einer Erklärung geeignet sein. Beispielsweise machte im Jahr 2015 eine Google-App Negativschlagzeilen, weil sie Fotos von Schwarzen Menschen als ‚Gorilla‘ klassifizierte (BBC, 2015). Obwohl diese Informationen wichtig sind, um das dahinter stehende Klassifikationssystem zu verbessern: Ein EXEMPLAR, das solche rassistischen Klassifikationskriterien visualisiert, könnte in manchen Situationen mehr schaden als nützen.¹³⁴ Andere EXEMPLARE sind unabhängig von den Adressat:innen ethisch und epistemisch inakzeptabel. Zum Beispiel haben Xiaolin Wu und Xi Zhang im Rahmen eines (inzwischen zurückgezogenen) Aufsatzes Fotos von verurteilten Straftäter:innen fusioniert. Dadurch hofften sie, „Subtypen krimineller Gesichter“ zu identifizieren (Wu & Zhang, 2016, S. 8; meine Übersetzung) – ein Vorgehen, das auf offensichtlich problematischen physiognomischen Annahmen beruht.¹³⁵

Viertens sind EXEMPLARE nicht zwangsläufig visuelle Repräsentationen, auch wenn ich mich in diesem Kapitel auf visuelle Erklärungen konzentriere. Dies lässt sich anhand von Elgins Diskussion der *Challenger*-Explosion aus dem Jahr 1986 illustrieren (siehe Abschnitt 7.1). Zur Erinnerung: Damals kamen mehrere Astronaut:innen ums Leben, weil mindestens ein Dichtungsring an der Raumfähre versagte. Elgin schildert, dass Ingenieur:innen der NASA den relevanten Entscheidungsträger:innen vor dem Start diverse Daten zugesandt hatten, die unter anderem auch Informationen über das Risiko eines solchen Vorfalls enthielten (Elgin, 2017, S. 262). Doch gingen die Informationen, mithilfe derer das Unglück hätte verhindert werden können, Elgin zufolge im Bericht unter. Analoge Fälle sind auch im KI-Kontext denkbar: Informationen über die unzureichende Fairness eines Empfehlungssystems könnten durch ein schriftliches EXEMPLAR hervorgehoben werden – etwa die explizite Ermahnung einer Watchdog-Organisation. Alternativ könnte diese Eigenschaft in einem schriftlichen SAMPLE untergehen – beispielsweise im offiziellen technischen Datenblatt des Empfehlungssystems, das Angaben zu dessen Genauigkeit bei Angehörigen verschiedener ethnischer Gruppen in einem langen, unübersichtlichen Textblock zusammenstellt. Das Grundprinzip ist universell anwendbar: Wenn man explizit macht, welche Eigenschaften ein Verstehensermöglicher exemplifiziert und welche Inter-

¹³⁴Danke an Dick Timmer, der mich auf diese Problematik hingewiesen hat.

¹³⁵Vielen Dank an Kevin Baum, der mich auf diesen bizarren Aufsatz aufmerksam gemacht hat.

pretationshilfen er bereitstellt, lassen sich Rückschlüsse auf das Verstehen ziehen, das er hervorbringen kann.

Im Anschluss an diese Klarstellungen können wir nun EXEMPLARE und ihren Zusammenhang zu instrumentellem Verstehen näher charakterisieren. In Abschnitt 12.2.1 arbeite ich fünf Eigenschaften heraus, die EXEMPLARE auszeichnen und besonders vorteilhaft für den Erwerb von (instrumentellem) Verstehen machen. Anschließend erläutere ich in Abschnitt 12.2.2, inwiefern EXEMPLARE die Erfüllung der Bedingungen instrumentellen Verstehens begünstigen.

12.2.1 Eigenschaften von EXEMPLAREN

Ich schlage vor, dass EXEMPLARE durch fünf Eigenschaften charakterisiert werden können: Sie (1) heben exemplifizierte Merkmale hervor, (2) spielen inerte Merkmale herunter, (3) haben einen klaren Geltungsbereich und exemplifizieren Merkmale, die sowohl (4) kontextuell relevant als auch (5) nachvollziehbar sind (Mann, 2025, S. 45).¹³⁶ Ich gehe nacheinander auf jeden dieser Vorzüge ein.

Hervorheben exemplifizierter Merkmale

EXEMPLARE heben exemplifizierte Merkmale hervor und grenzen sie dadurch von inerten Merkmalen ab. Das Aufzeigen exemplifizierter Merkmale ist eine Interpretationshilfe, die Interpret:innen dabei unterstützt, die vom EXEMPLAR typisierte Extension zu bestimmen. Wie in Abschnitt 12.1 dargestellt, kann die Hervorhebung exemplifizierter Merkmale auf unterschiedlichen Wegen erreicht werden, etwa durch zusätzliche Erläuterungen, die Ergänzung relevanter Kontraste oder das Entfernen ablenkender Merkmale (ebd.).

Betrachten wir ein Beispiel für die gelungene Hervorhebung relevanter Merkmale aus dem XAI-Kontext: Ein Problem von Bildklassifikationssystemen besteht darin, dass sie sich mitunter auf problematische Proxy-Merkmale stützen (siehe Kapitel 1). Solche Proxies sind nicht generalisierbar, besitzen also keine gute Vorhersagekraft in Bezug auf neue Daten. Sebastian Lapuschkin et al. fanden mithilfe ihres Erklärbarkeitsansatzes SpRAY heraus, dass ein Bildklassifikationssystem bestimmte Fotos aufgrund eines Wasserzeichens am Bildrand der Bildklasse ‚Pferd‘ zugewiesen hatte (Lapuschkin et al., 2019). Der Grund:

¹³⁶ Ähnliche Vorschläge finden sich in unterschiedlichen Disziplinen und philosophischen Subdebatten wieder. Andrés Páez und Brent Mittelstadt et al. betonen wie ich die Wichtigkeit (und Schwierigkeit) eines klaren Geltungsbereichs von XAI-Erklärungen (Mittelstadt et al., 2019, S. 281–282; Páez, 2019, S. 450). Julie Jebeile hebt als Vorteil visueller Repräsentationen unter anderem hervor, dass sie relevante Aspekte aufdecken und selektiv hervorheben können (Jebeile, 2018b, Kap. 4). Im Kontext der Erziehungspsychologie formulieren Matthew McCrudden und David Rapp Designempfehlungen für visuelle Darstellungen, die der kognitiven Verarbeitung der enthaltenen Informationen zuträglich sein und dadurch das Lernen erleichtern sollen. Dazu zählt auch, Betrachter:innen bei der Identifikation relevanter Informationen z. B. durch Hervorhebungen zu unterstützen und irrelevante Informationen möglichst wegzulassen (McCrudden & Rapp, 2017). Auch die in Abschnitt 7.2 geschilderten Vorteile von Idealisierungen, die Will Fleisher (2022) zufolge auch für XAI-Kontexte relevant sind, ähneln einigen Eigenschaften von EXEMPLAREN.

In den Trainingsdaten waren viele Pferdefotos eines bestimmten Fotografen enthalten, der seinen Fotos ein Wasserzeichen hinzugefügt hatte. Natürlich ist dieses Wasserzeichen kein charakteristisches Merkmal von Pferdebildern, sondern ein Proxy, der jenseits der Trainingsdaten so gut wie keine Vorhersagekraft besitzt. Aus menschlicher Sicht sind Proxies wie diese häufig kontraintuitiv und daher leicht zu übersehen. Ein EXEMPLAR, das Proxy-Merkmale hervorhebt, kann solche Probleme manifest machen und dabei helfen, Bugs oder problematische Verzerrungen in KI-Systemen rechtzeitig zu bemerken (Mann, 2025, S. 45).

Herunterspielen inerter Merkmale

EXEMPLARE spielen inerte Merkmale herunter, um sie von exemplifizierten Merkmalen abzugrenzen und so die Interpretation zu erleichtern. Das Herunterspielen inerter Merkmale und die Hervorhebung exemplifizierter Merkmale gehen oft Hand in Hand. Daher können die oben erwähnten Strategien zur Hervorhebung exemplifizierter Eigenschaften auch als Strategien für das Herunterspielen inerter Merkmale angesehen werden (ebd., S. 46).

Oftmals ist das Weglassen von irrelevanten Merkmalen die effektivste Möglichkeit, um sie herunterzuspielen. Bei visuellen EXEMPLAREN ist dies aufgrund der expressiven Einschränkungen visueller Repräsentationen aber nicht immer möglich (siehe Abschnitt 7.3.2). Beispielsweise lässt sich die Instanziierung von Farbe nicht vermeiden (Elgin, 2017, S. 192). Ein Bild, das eine ‚farblose‘ Form zeigt, ist nicht von einem Bild zu unterscheiden, das z. B. eine weiße Form zeigt. In diesen Fällen sind zusätzliche Interpretationshilfen erforderlich, etwa in Form von linguistischen Ergänzungen.¹³⁷ Andere visuelle Merkmale können nur durch die zusätzliche Instanziierung weiterer Eigenschaften exemplifiziert werden. Beispielsweise kann das Vorhandensein eines Vogelschnabels nur mithilfe der Instanziierung einer bestimmten Schnabelposition exemplifiziert werden – auch, wenn die Schnabelposition selbst irrelevant ist (Mann, 2025, S. 46). Ich komme in Abschnitt 12.4.2 auf diese Problematik zurück.

Demgegenüber haben linguistische Verstehensermöglicher (z. B. die verbale Erklärung einer Expertin) den Vorteil, viele Eigenschaften und deren Relationen untereinander präziser und expliziter ausdrücken zu können. Sprachliche Verstehensermöglicher können exemplifizierte Aspekte dadurch leichter von inerten Merkmalen abgrenzen (etwa durch Negationen). Gleichzeitig legt der oben beschriebene Fall der *Challenger*-Explosion nahe, dass das linguistische Format nicht garantiert, dass es sich bei einem Verstehensermöglicher um ein EXEMPLAR im hier beschriebenen Sinne handelt: Wenn relevante Informationen in einem ausufernden Text untergehen, werden inerte bzw. exemplifizierte Merkmale nicht als solche erkennbar.

¹³⁷Solche Überlegungen finden sich auch in der informatischen XAI-Literatur, siehe Fußnote 86.

Klarer Geltungsbereich

Der Geltungsbereich von EXEMPLAREN ist so eindeutig wie möglich. Die beiden vorangegangenen Eigenschaften von EXEMPLAREN erleichtern die Interpretation, indem sie dabei helfen, exemplifizierte bzw. inerte Merkmale zu erkennen und dadurch die Gemeinsamkeiten der Mitglieder der typisierten Extension zu identifizieren (z. B. ein rotes Brustgefieder). Im Gegensatz dazu hilft ein klarer Geltungsbereich dabei zu bestimmen, um was für eine Extension es sich handelt – beispielsweise um die Gruppe der typischen Rotkehlchen (ebd.).

Eine Möglichkeit, um den Geltungsbereich eines EXEMPLARS deutlich zu machen, können Informationen über das KI-System und seine Grenzen sein. Nehmen wir ein Klassifikationssystem zur Erkennung von Hautkrebs, das hauptsächlich mit Daten von hellhäutigen Menschen trainiert wurde (Wen et al., 2021). In einem solchen Szenario sollten die Interpret:innen wissen, dass ein EXEMPLAR der Bildklasse ‚Melanom‘ nicht unbedingt für Daten von Menschen mit dunklerer Haut gültig ist. Auch der XAI-Ansatz, der für die Generierung des EXEMPLARS verwendet wurde, kann sich auf dessen Geltungsbereich auswirken (Mann, 2025, S. 46). Es ist beispielsweise wichtig zu wissen, dass lokale Erklärungen einen viel kleineren Teil der Eingabedaten abdecken als globale Erklärungen (siehe Abschnitt 3.2.2).

Schließlich beeinflussen auch die zuvor beschriebenen Eigenschaften von EXEMPLAREN ihren Geltungsbereich. EXEMPLARE bieten epistemischen Zugang zu potenziell komplexen Referenten (z. B. den Merkmalen, die tiefe neuronale Netze während des Trainings erlernt haben), indem sie einige ihrer Eigenschaften in den Vordergrund rücken und andere herunterspielen. Das kann dazu führen, dass EXEMPLARE Falschheiten beinhalten. In der Einleitung dieses Kapitels habe ich vorgeschlagen, dass solche Falschheiten den Verstehenswerb nicht notwendigerweise behindern und im Gegenteil sogar vorteilhaft sein können. Doch das ist nur dann der Fall, wenn die EXEMPLARE korrekt interpretiert werden.

Stellen wir uns ein EXEMPLAR vor, das Merkmale exemplifiziert, die ein KI-System nutzt, um Fotos der Bildklasse ‚Rotkehlchen‘ zuzuweisen. Da moderne Bildklassifikationssysteme eine Vielzahl von Bildeigenschaften heranziehen, ist anzunehmen, dass dieses EXEMPLAR nur eine Auswahl der einflussreichen Merkmale exemplifizieren kann. Es verschafft einer Interpretin daher auch keinen epistemischen Zugang zu sämtlichen Merkmalen von Bildern dieser Klasse. Angesichts dessen ist es wichtig, sicherzustellen, dass Interpret:innen anhand von EXEMPLAREN nur auf Instanzen jener Extension schließen, die das EXEMPLAR (näherungsweise) korrekt repräsentiert – also z. B. die Gruppe der typischen Rotkehlchenbilder. Dafür muss wiederum der Geltungsbereich des Beispiels bekannt sein (ebd.).

Exemplifikation kontextuell relevanter Merkmale

EXEMPLARE exemplifizieren kontextuell relevante Merkmale. Ein Bild, das charakteristische Eigenschaften von Rotkehlchen hervorhebt, mag für Arda und Moira aus den eingangs beschriebenen Fallbeispielen als EXEMPLAR durchgehen – für eine NASA-Mitarbeiterin, die den Absturz der *Challenger*-Raumfähre untersucht, ist ein solches Foto hingegen

irrelevant. Das Kriterium der Kontextsensitivität von EXEMPLAREN deckt sich mit der gängigen Annahme aus der XAI-Literatur, dass der Kontext die Eignung einer gegebenen Erklärung beeinflusst (siehe Abschnitt 4.2.3). Viele existierende Erklärbarkeitsansätze für bildverarbeitende Systeme bilden diese kontextuell wechselnden Anforderungen ab.

Eine XAI-Methode, die sich an die Bedürfnisse der Adressatin einer Erklärung anpassen lässt, ist beispielsweise TCAV (*Testing with Concept Activation Vectors*; B. Kim et al., 2018). Mit TCAV kann der Einfluss von nutzergenerierten ‚Begriffen‘ (z. B. *gestreift*) auf die finale Klassifikation (z. B. ‚Zebra‘) überprüft werden, um kontextspezifische Fragen zur Klassifikation eines Bildes zu beantworten. Mithilfe der unterschiedlichen Farben mancher Salienzkarten kann der relative Einfluss verschiedener Bildregionen exemplifiziert werden (siehe etwa Abbildungen 3.2b und 7.1). Kontrafaktische Abhängigkeiten lassen sich durch visuelle kontrafaktische Erklärungen exemplifizieren (z. B. White et al., 2023). Und mehrere EXEMPLARE können verschiedene ‚Klassifikationsstrategien‘ eines Systems visualisieren – Lapuschkin et al. haben mithilfe mehrerer Salienzkarten beispielsweise unterschiedliche Strategien eines KI-Systems für die Klassifikation von Pferdebildern sichtbar gemacht (etwa die Detektion von Hürden oder einer Reiterin, siehe Lapuschkin et al., 2019, Abb. 3 auf S. 5). Schließlich können zusätzliche linguistische Erklärungen verwendet werden, um nicht visualisierbare Klassifikationskriterien zu beschreiben und so die expressiven Einschränkungen visueller Repräsentationen zu kompensieren, die ich in Abschnitt 7.3.2 beschrieben habe (Mann, 2025, S. 46).

Exemplifikation nachvollziehbarer Merkmale

Die von EXEMPLAREN exemplifizierten Merkmale müssen für die menschliche Interpretin nachvollziehbar sein in dem Sinne, dass sie auf Eigenschaften anderer Instanzen übertragen werden können. Andernfalls würde das EXEMPLAR nicht ‚über sich hinausweisen‘. Um beispielsweise zu verstehen, wie spezifische Bildmerkmale und eine bestimmte Bildklasse zusammenhängen, müssen diese Merkmale aus menschlicher Sicht Sinn ergeben. Diese Merkmale können dann durch ein EXEMPLAR exemplifiziert werden (ebd., S. 46–47).

Merkmale, die in diesem Sinne nachvollziehbar sind, müssen keine Entsprechung in existierenden menschlichen Begriffen haben, und sie müssen nicht als distinkte Merkmale erkannt worden sein, bevor sie exemplifiziert werden (vgl. Elgin, 2017, S. 187). Beispielsweise haben Ludwig Schubert et al. durch eine Kombination verschiedener XAI-Ansätze ‚Hoch-niedrig-Frequenz-Detektoren‘ identifiziert. Diese Neuronen reagieren auf Bildbereiche, die einen Übergang zwischen einer hohen und niedrigen Ortsfrequenz¹³⁸ aufweisen (Schubert et al., 2021). Die Ortsfrequenz ist kein Merkmal, das Menschen bewusst nutzen, um Objekte oder Bilder zu kategorisieren, aber es ist ein Merkmal, das Menschen nachvollziehen können. Wenn die exemplifizierten Merkmale hingegen zu fremdartig sind, um

¹³⁸Die mathematischen Hintergründe beiseite lassend, sind Bildbereiche mit feinen Texturen und Kanten (beispielsweise ein Foto von Gesteinsschotter) üblicherweise hochfrequent und solche mit großflächigen Farben oder allmählichen Farbverläufen niedrigfrequent (etwa der blaue Himmel auf einem Landschaftsbild).

nachvollziehbar zu sein, scheint Verstehen unmöglich – nicht umsonst handelt es sich bei fremdartigen Repräsentationen um eine in KI-Kontexten wichtige Opazitätsquelle (siehe Abschnitt 2.3.1). Diese Problematik läuft letztlich auf die Frage hinaus, ob moderne Bildklassifikationssysteme im Besonderen und tiefe neuronale Netze im Allgemeinen überhaupt verständlich gemacht werden können. Ich diskutiere diesen Aspekt in Abschnitt 12.4.1.

12.2.2 Instrumentelles Verstehen durch EXEMPLARE

Betrachten wir abschließend, wie sich EXEMPLARE auf die Bedingungen instrumentellen Verstehens auswirken. Elgin zufolge kann ein Beispiel, das Merkmale seines Referenten exemplifiziert, Verstehen dieses Referenten hervorbringen, weil es dessen kontextuell relevante Merkmale salient macht (Elgin, 2017, S. 249). Elgin ergänzt diesen Vorschlag durch ihre Theorie des Verstehens, die maßgeblich auf der Lockerung des Faktizitätskriteriums (siehe Kapitel 8) und dem Überlegungsgleichgewicht (siehe Abschnitt 10.2) beruht. Doch auch anhand meiner Definition instrumentellen Verstehens lässt sich aufzeigen, wie effektive Beispiele Verstehen durch Exemplifikation herbeiführen können. Hier ist noch einmal mein Definitionsvorschlag aus Abschnitt 11.1.1:

Instrumentelles Verstehen: Für Kontext K – bestehend aus (i) einem Stakeholder S , (ii) der Rolle R von S , (iii) dem Desideratum D , das S in Bezug auf ein KI-System und/oder das übergeordnete soziotechnische System besitzt sowie (iv) den mit der (Nicht-)Erfüllung von D verbundenen Chancen und Risiken CR – gilt:

S versteht instrumentell genau dann, wenn die folgenden Bedingungen erfüllt sind:

- (1) S besitzt ein Überzeugungssystem, das die folgenden Bedingungen erfüllt: (*Überzeugungsbedingung und Komplexitätsbedingung*)
 - (a) Die Inhalte der darin enthaltenen Überzeugungen sind Propositionen über systembezogene Aspekte. (*Objektbedingung*)
 - (b) Die darin enthaltenen zentralen, d. h. für die Erfüllung von D relevanten und ggf. gemäß R eingegrenzten, Überzeugungen sind zumindest annähernd wahr. (*Faktizitätsbedingung*)
 - (c) Die darin enthaltenen Überzeugungen sind *ultima facie* gerechtfertigt, wobei der dafür erforderliche Grad der Rechtfertigung von CR abhängt. (*Rechtfertigungsbedingung*)
- (2) S verfügt über für die Erfüllung von D relevante und ggf. gemäß R eingegrenzte verstehensbezogene Fähigkeiten. (*Fähigkeitsbedingung*)

Betrachten wir genauer, weshalb EXEMPLARE dem Erwerb instrumentellen Verstehens zuträglich sind, und beginnen mit Bedingung (1a). Dass EXEMPLARE einem Stakeholder dabei helfen können, Überzeugungen über systembezogene Aspekte zu erwerben, dürfte im Lichte der bisherigen Diskussion auf der Hand liegen: EXEMPLARE heben selektiv bestimmte

Eigenschaften ihres Referenten hervor, typisieren eine bestimmte Extension von Instanzen und können einem Stakeholder dadurch epistemischen Zugang zu systembezogenen Aspekten verschaffen. Im Kontext der Bildklassifikation zählen dazu beispielsweise die ausschlaggebenden Merkmale für die Klassifikation eines Bildes oder typische Eigenschaften von Mitgliedern einer ganzen Bildklasse. Kurz: Indem sie bestimmte Aspekte salient machen, können EXEMPLARE zu einem Überzeugungssystem über systembezogene Aspekte beitragen und dadurch die Erfüllung von Bedingung (1a) instrumentellen Verstehens unterstützen.

Kommen wir zu Bedingung (1b). EXEMPLARE exemplifizieren kontextuell relevante Merkmale und besitzen einen klaren Geltungsbereich. Dadurch tragen sie dazu bei, dass die für die Erfüllung des jeweiligen Desideratums zentralen Überzeugungen, die ein Stakeholder auf Basis eines solchen Verstehensermöglichers erwirbt, wahr oder zumindest annähernd wahr sind. Auf dieser Grundlage könnte ein Stakeholder beispielsweise abschätzen, ob ein Bildklassifikationssystem vertrauenswürdig und robust ist oder problematische Verzerrungen aufweist.

Wie in Abschnitt 12.2.1 dargestellt, sind EXEMPLARE normalerweise keine vollständig akkuraten Repräsentationen ihres Referenten. Sie können diverse ‚gelungene Falschheiten‘ aufweisen, die potenziell zu falschen Überzeugungen über systembezogene Aspekte führen. Da ein EXEMPLAR *per definitionem* zu den kontextuellen Anforderungen des fraglichen Stakeholders passt, beeinträchtigen diese falschen Überzeugungen aber nicht den Erwerb instrumentellen Verstehens. Im Gegenteil: Weil ein EXEMPLAR das Formen wahrer, kontextuell zentraler Überzeugungen unterstützt (indem es relevante Merkmale seines Referenten exemplifiziert), hilft es dem Stakeholder dabei, das moderat faktive Überzeugungssystem über systembezogene Aspekte auszubilden, welches laut Bedingung (1b) erforderlich für instrumentelles Verstehen ist.

Subbedingung (1c) betrifft die Rechtfertigung der Überzeugungen, die ein Stakeholder z. B. mithilfe eines Verstehensermöglichers erwirbt. Dieser Aspekt instrumentellen Verstehens wird durch meine Charakterisierung von EXEMPLAREN nicht direkt adressiert, ich möchte deswegen nur einige Anmerkungen dazu ergänzen. Um Überzeugungen rechtfertigen zu können, sollten EXEMPLARE z. B. durch einen zuverlässigen Erklärbarkeitsansatz generiert werden, der im gegebenen Kontext aussagekräftige und hinreichend akkurate Verstehensermöglicher produziert. Aufgrund der fundamentalistischen Dimension der Rechtfertigung müssen die resultierenden Überzeugungen letztlich durch basale Überzeugungen gerechtfertigt sein. Aufgrund der kohärentistischen Dimension müssen sie mit den anderen Überzeugungen des Überzeugungssystems kohärieren, also in angemessenen explanatorischen, logischen oder probabilistischen Stützungsrelationen stehen und nicht zu Widersprüchen führen. Ob ein EXEMPLAR Bedingung (1c) von instrumentellem Verstehen zuträglich ist, hängt also auch davon ab, welches Vorwissen bzw. ‚Vorverstehen‘ der Stakeholder besitzt und wie das EXEMPLAR ausgewählt oder erzeugt wurde.

Bleibt noch Bedingung (2). Das Überzeugungssystem, das bestenfalls aus EXEMPLAREN resultiert, bildet die Grundlage für verstehensbezogene Fähigkeiten. Auch Elgin ist der

Ansicht, dass Verstehen mit Fähigkeiten einhergeht: „[understanding] involves being able to draw inferences, raise questions, frame potentially fruitful inquiries, and so forth“ (Elgin, 2017, S. 3). Dass EXEMPLARE verstehensbezogene Fähigkeiten hervorbringen können, lässt sich exemplarisch an zwei der am häufigsten diskutierten verstehensbezogenen Fähigkeiten verdeutlichen: dem kontrafaktischen Schließen und dem Ziehen von Schlussfolgerungen über ähnliche Fälle (siehe Tabelle 9.4). Wir erinnern uns, dass ein Beispiel die Extension jener Instanzen typisiert, welche die exemplifizierten Merkmale teilen. Das Erfassen dieser Extension ermöglicht es zu beurteilen, ob eine bestimmte Instanz zu dieser Extension gehört. Entsprechend kann ein EXEMPLAR der Interpretin die Fähigkeit verleihen, Schlussfolgerungen über Instanzen zu ziehen, die vom Beispiel abweichen.

Denken wir noch einmal an die ornithologisch ungebildete Moira aus *Schnappschuss* zurück, die nicht weiß, wie Rotkehlchen aussehen. Nehmen wir an, dass Moira eine erfahrene Vogelbeobachterin zurate zieht, die ihr ein Beispiel für einen solchen Vogel zeigt, und dass dessen rötliches Brust- und Kopfgefieder die exemplifizierten Merkmale darstellen. Entsprechend typisiert der Vogel die Extension jener Vögel, die diese Eigenschaften aufweisen. Wenn Moira das Beispiel korrekt interpretiert, besitzt sie die Fähigkeit, Schlüsse wie die folgenden zu ziehen: *Wenn der Vogel kein rötliches Brustgefieder besäße, wäre es kein Rotkehlchen. Der Kopf dieses anderen Vogels ist ausschließlich braun, also handelt es sich nicht um ein Rotkehlchen. Dieser Vogel hingegen weist ein rötliches Kopf- und Brustgefieder auf, entsprechend ist es ebenfalls ein Rotkehlchen.* Solche Fähigkeiten erscheinen auf den ersten Blick banal. Doch stellen wir uns vor, dass jemand analoge Fähigkeiten in Bezug auf ein Bildklassifikationssystem besitzt: *Dieses Bild weist Merkmal X auf, also wird es (wahrscheinlich) als k klassifiziert. Wenn das Bild X nicht besäße, würde es (wahrscheinlich) nicht als k klassifiziert. Dieses andere Bild besitzt ebenfalls X, folglich wird es vom System (wahrscheinlich) als k kategorisiert.* Aufgrund der komplexen Klassifikationskriterien moderner Bildklassifikatoren können solche verstehensbezogenen Fähigkeiten eine signifikante epistemische Errungenschaft darstellen. Da ein EXEMPLAR einfacher zu interpretieren ist als ein SAMPLE, begünstigt es den Erwerb solcher Fähigkeiten. Diese können wiederum erforderlich sein, um Bedingung (2) instrumentellen Verstehens zu erfüllen.

12.3 Generieren XAI-Methoden EXEMPLARE?

Wir wissen nun, wodurch sich EXEMPLARE auszeichnen, und wie sie instrumentelles Verstehen hervorbringen können. In diesem Abschnitt vergleiche ich die bislang nur theoretisch motivierten EXEMPLARE mit den visuellen Erklärungen existierender XAI-Methoden, die zur Erklärung von Bildklassifikationssystemen verwendet werden. Ich möchte illustrieren, dass meine Charakterisierung von EXEMPLAREN dabei helfen kann, bestehende Verstehensermöglicher zu evaluieren und womöglich neue Ansätze zu inspirieren, die instrumentellem Verstehen zuträglich sind (Mann, 2025, S. 47). Dadurch trägt dieser Vergleich indirekt auch zur Erfüllung des Adäquatheitskriteriums der Fruchtbarkeit bei, das ich im Rahmen meiner

Explikation instrumentellen Verstehens formuliert habe (siehe Abschnitte 4.3.2 und 11.2.2).

Welche systembezogenen Aspekte ein gegebener Stakeholder verstehen muss, um instrumentelles Verstehen zu besitzen, hängt von kontextuellen Faktoren ab (siehe Abschnitte 6.3 und 11.1). Bezogen auf den Bereich der Bildklassifikation muss jemand vielleicht die grundsätzlichen technischen Hintergründe dieses Verfahrens verstehen, warum ein einzelnes Bild in einer bestimmten Weise klassifiziert wurde oder in welchem Teil des Systems ein spezifisches Bildmerkmal detektiert wurde. Um die nachfolgende Diskussion zu vereinfachen, konzentriere ich mich hier auf ein einzelnes epistemisches Ziel: verstehen, warum Bilder von einem Klassifikationssystem einer bestimmten Bildklasse zugeordnet werden (z. B. verstehen, warum Bilder vom System als ‚Rotkehlchen‘ klassifiziert werden). Gemeint ist also nicht das Verständnis einer einzelnen Klassifikation, sondern jenes einer ganzen Bildklasse – im Sinne der in Abschnitt 3.2.2 vorgenommenen Systematisierung von XAI-Methoden anhand ihres Geltungsbereichs handelt es sich gewissermaßen um ein ‚globales‘ Verständnis. Ich bezeichne dieses spezifische epistemische Ziel als ‚Verstehen_{Klasse}‘ (Mann, 2025, S. 38). Ein solches Verständnis kann einem Stakeholder z. B. ermöglichen, die Vertrauenswürdigkeit und Robustheit eines Klassifikationssystems zu beurteilen oder Fehler zu beseitigen. Nachfolgend lasse ich der Übersichtlichkeit halber aber das konkrete Desideratum des Stakeholders außer Acht (dasselbe gilt für seine Rolle und die kontextuellen Chancen bzw. Risiken). Für die Zwecke dieses Abschnitts genügt festzuhalten, dass Verstehen_{Klasse} für bestimmte Instanzen instrumentellen Verstehens von Belang sein kann.

Welche systembezogenen Aspekte sind also für Verstehen_{Klasse} relevant? Ein plausibler Kandidat sind jene Merkmale, die das Klassifikationssystem nutzt, um Bilder der fraglichen Bildklasse zuzuweisen. Diese Merkmale sollte die Erklärung exemplifizieren, um als EXEMPLAR in Frage zu kommen. Wenn ein System Bilder mit Pixeln, die einem rötlichen Brustgefieder entsprechen, als Mitglieder der Klasse ‚Rotkehlchen‘ klassifiziert, muss das Beispiel dieses Merkmal exemplifizieren. Zwar stützen sich die meisten modernen Bildklassifikationssysteme auf eine zu große Anzahl von Bildmerkmalen, als dass diese noch vollständig kognitiv erfasst (oder exemplifiziert) werden könnten. Doch zumindest können ein oder mehrere EXEMPLARE eine Auswahl der kontextuell relevanten, z. B. der einflussreichsten, Merkmale exemplifizieren.

Es gibt eine Vielzahl von XAI-Ansätzen, die darauf abzielen, die Klassifikation von Bildern durch visuelle Verstehensermöglicher zu erklären (eine Übersicht bietet z. B. Molnar, 2022). In den folgenden Abschnitten stelle ich eine Auswahl dieser Erklärbarkeitsmethoden vor. Diese Auswahl ist nicht auf Methoden beschränkt, die in Taxonomien von XAI-Methoden gängigerweise als beispielbasierte Erklärungen kategorisiert werden (siehe etwa Arrieta et al., 2020; Belle & Papantonis, 2021; Z. C. Lipton, 2018; Schwalbe & Finzel, 2024; Speith, 2022). Auch Erklärungen, die mit anderen gängigen XAI-Ansätzen erstellt werden, können Verstehen hervorbringen, indem sie einige ihrer Merkmale exemplifizieren.

Ich habe die von mir berücksichtigten Methoden in drei Gruppen eingeteilt: erstens Methoden, die mehrere Instanzen einer Bildklasse anführen (Abschnitt 12.3.1); zweitens

Ansätze, die einflussreiche Bereiche innerhalb einer Instanz der Bildklasse hervorheben (Abschnitt 12.3.2) und drittens Methoden, die ein künstliches Bild erzeugen, indem sie für die Bildklasse charakteristische Merkmale visualisieren (Abschnitt 12.3.3).

Meine Leitfrage ist, ob die resultierenden Erklärungen als EXEMPLARE für den Erwerb von Verstehen_{Klasse} gelten können – konkret, ob die Erklärungen geeignet sind, um dafür erforderliche Überzeugungen und Fähigkeiten hervorzubringen. Gemäß meiner Charakterisierung von EXEMPLAREN in Abschnitt 12.2.1 sollten die visuellen Erklärungen dafür exemplifizierte Merkmale hervorheben, inerte Merkmale herunterspielen, einen klaren Geltungsbereich aufweisen und Merkmale exemplifizieren, die sowohl kontextuell relevant als auch nachvollziehbar sind.

12.3.1 Gruppe 1: Ansätze, die Instanzen einer Bildklasse anführen

Der naheliegendste Weg, um ein Beispiel zu erhalten, besteht darin, (irgend)eine Instanz auszuwählen, welche die zu exemplifizierenden Merkmale instanziiert. Einen solchen Ansatz verfolgen die Erklärbarkeitsmethoden der ersten Gruppe. In der Regel liefern diese Ansätze nicht nur eines, sondern mehrere Beispiele einer Bildklasse. Dadurch kann eine größere Anzahl von charakteristischen Merkmalen exemplifiziert und die Komplexität der typischen Merkmale von Mitgliedern einer Bildklasse besser abgebildet werden. Wie sich zeigen wird, ist diese Strategie jedoch nicht ausreichend, um eine korrekte Interpretation der Beispiele zu gewährleisten. Entsprechend werde ich argumentieren, dass es sich bei den Erklärungen der ersten Gruppe nicht um EXEMPLARE handelt (Mann, 2025, S. 42).

Betrachten wir also einige XAI-Ansätze dieser Gruppe genauer. Eine von Jacob Bien und Robert Tibshirani vorgeschlagene Methode wählt so genannte ‚Prototypen‘ aus. Ein Prototypensatz besteht aus mehreren Dateninstanzen, die repräsentativ für verschiedene Cluster in der Datenverteilung einer Bildklasse sind und die deren Elemente daher visuell zusammenfassen können (Bien & Tibshirani, 2011, S. 2418). Der MMD-Critic-Ansatz von Been Kim et al. ergänzt diese Prototypen um ‚Kritiken‘ (engl. *criticisms*). Kritiken sind Instanzen, die Teil der Datenverteilung sind, aber durch Prototypen nicht gut repräsentiert werden. In Kombination mit den Prototypen sollen die Kritiken die Vielfalt des Datensatzes besser widerspiegeln (B. Kim et al., 2016; siehe Abbildung 12.2). Eine weitere Möglichkeit, typische Instanzen einer Bildklasse zu identifizieren, besteht darin, ‚einflussreiche Beispiele‘ bzw. ‚Datensatz-Beispiele‘ auszuwählen – also Instanzen der Trainingsdaten, die einen großen Einfluss auf die Modellparameter hatten bzw. die eine bestimmte Einheit eines neuronalen Netzes (z. B. ein bestimmtes Neuron) maximal aktivieren (siehe dazu Abschnitt 3.2.2).

Damit liefern diese Methoden mehrere Beispiele, die Merkmale salient machen sollen, auf die sich ein gegebenes Bildklassifikationssystem bei der Klassifizierung von Bildern einer bestimmten Klasse typischerweise stützt – es geht also um Merkmale, die für Verstehen_{Klasse} relevant sind. Doch damit ein Stakeholder dieses Verstehen erwerben kann, müssen die Beispiele auch richtig interpretiert werden.



Abbildung 12.2: MMD-Critic-Erklärung für zwei Bildklassen, jeweils bestehend aus einem Prototypensatz und einer Gruppe von Kritiken. Bildquelle: B. Kim et al. (2016). Siehe Teil IV für Details.

Wir erinnern uns, dass es zwei Formen der Interpretation gibt: Zum einen muss erfasst werden, worauf die Extension hinausläuft, die ein Beispiel typisiert (also z. B., ob es sich um die Extension der typischen Rotkehlchenbilder handelt). Der Geltungsbereich der Erklärungen dieser Methoden ist vergleichsweise klar – es handelt sich um Beispiele, welche die Extension der typischen Mitglieder einer Bildklasse typisieren sollen (Mann, 2025, S. 42).

Zum anderen geht es bei der Interpretation eines Beispiels darum, die exemplifizierten Merkmale zu identifizieren und von inerten bzw. stützenden Merkmalen abzugrenzen. Die XAI-Ansätze der ersten Gruppe bieten jedoch nur wenige Interpretationshilfen, die eine Interpretin dabei unterstützen. Als Beispiele dienen unveränderte Dateninstanzen, die eine Fülle von Merkmalen instanziierten können. Die Interpretin muss entscheiden, welche der Merkmale exemplifizierte Merkmale sind. Da Bildklassifikationssysteme häufig andere Muster erkennen als Menschen, ist dies keineswegs trivial (vgl. Abschnitt 12.2). Ein Mensch, der mit einer Reihe von Rotkehlchenbildern konfrontiert wird, greift auf sein Weltwissen zurück und nimmt wahrscheinlich an, dass bestimmte Merkmale der Vogelart durch die Beispiele exemplifiziert werden. Es kann jedoch sein, dass das Modell eine Repräsentation der Bildklasse erlernt hat, die nicht mit der menschlichen Vorstellung von Rotkehlchen übereinstimmt (ebd.). Vielleicht reagiert das System stattdessen auf Wasserzeichen, Aspekte des Hintergrunds oder andere Proxies, die ein Mensch womöglich nicht einmal bemerkt (siehe Abschnitt 12.2.1). Wenn man zum Beispiel die rechte obere Bildreihe in Abbildung 12.2 betrachtet, könnte das dadurch exemplifizierte Merkmal ebenso gut der jeweils im Hintergrund sichtbare grüne Rasen sein – anstelle von charakteristischen Eigenschaften der Hunderasse, wie man intuitiv annehmen würde.

Wenn es der Interpretin nicht gelingt, die exemplifizierten Merkmale zu identifizieren, erlangt sie auch kein $\text{Verstehen}_{\text{Klasse}}$: Zum einen erwirbt sie keine wahren Überzeugungen darüber, welche Merkmale das System mit der fraglichen Bildklasse typischerweise assoziiert. Zum anderen verfügt sie mangels eines geeigneten Überzeugungssystems auch

nicht über verstehensbezogene Fähigkeiten. Sie kann beispielsweise nicht darauf schließen, ob ein gegebenes Bild zu der Extension gehört, die durch die Beispiele typisiert wird, oder erklären, warum bestimmte Bilder der fraglichen Klasse zugewiesen werden (ebd.).

Zusammenfassend lässt sich festhalten, dass die Erklärungen von Methoden der ersten Gruppe vergleichsweise ambig sind und kaum Interpretationshilfen bieten, um diese Ambiguität zu reduzieren. Sie heben exemplifizierte Merkmale nicht ausreichend hervor und spielen inerte Merkmale nur ungenügend herunter – nicht zuletzt, weil diese Erklärungen sehr viele inerte Eigenschaften instanziierten, die leicht mit exemplifizierten Merkmalen verwechselt werden können. In dieser Hinsicht ähneln sie dem zufällig ausgewählten Rotkehlchenfoto voller ‚Klimbim‘, das Moira aus *Schnappschuss* betrachtet. Das macht die Erklärungen dieser Ansätze zu SAMPLES. Natürlich können diese Methoden in vielen Fällen ihren Zweck erfüllen. Mein Punkt ist lediglich, dass die resultierenden Beispiele zumindest in Hinblick auf Verstehen_{Klasse} verbesserungswürdig sind (ebd., S. 43).

12.3.2 Gruppe 2: Ansätze, die einflussreiche Bereiche in einer Instanz der Bildklasse hervorheben

Lokale Attributionsmethoden (siehe Abschnitt 3.2.1) bilden wahrscheinlich die bekannteste Familie von Erklärbarkeitsansätzen im Bereich der Bildklassifizierung. Zur Erinnerung: Solche Methoden heben diejenigen Bereiche eines Eingabebildes hervor, die für die Klassifizierung am relevantesten waren – zumeist in Form von Salienzkarten. Diese Salienzkarten fügen häufig eine farbige Bildebene in das ursprüngliche Bild ein (wie in Abbildung 12.3) oder maskieren weniger einflussreiche Bereiche (wie in Abbildung 3.3b). Übertragen auf Exemplifikation heißt das, dass diese Methoden stützende Merkmale (d. h. farbige oder maskierte Bereiche) in das Bild integrieren. Dadurch versuchen Salienzkarten zu verdeutlichen, welche Bildbereiche exemplifizierte bzw. inerte Merkmale instanziierten (ebd.).

Als lokale Methoden sind Attributionsansätze nicht darauf ausgelegt, Verstehen_{Klasse} hervorzubringen. Sie zielen nicht darauf ab, die Klassifikation einer größeren Gruppe von Bildern zu erklären, sondern lediglich einer einzelnen Instanz. Eine Salienzkarte typisiert daher nur eine kleine Extension. Deren genauer Umfang bleibt in der Regel unklar, weil der Geltungsbereich einer lokalen Erklärung schwierig zu bestimmen ist (Belle & Papantonis, 2021, S. 13; Ribeiro et al., 2018, S. 1528). Dies erschwert die korrekte Interpretation: Es bleibt offen, worin die Extension besteht, die lokale Methoden typisieren. Für den Erwerb von Verstehen_{Klasse} bedeutet dies, dass lokale Salienzkarten der Interpretin nicht zu der Fähigkeit verhelfen, Schlüsse über andere Instanzen zu ziehen, da unklar ist, wie groß der Geltungsbereich der Erklärung ist.

Der Erklärbarkeitsansatz ‚Anchors‘ adressiert diese Schwierigkeit (ebd.). Anchors erzeugt lokale Erklärungen, die aus Wenn-dann-Regeln bestehen. Diese Regeln ‚verankern‘ die Klassifikation, da sie für eine bestimmte Gruppe von Instanzen gelten. Ein Anker besteht aus einem oder mehreren Feature-Werten. Wenn der Anker in einer bestimmten Dateninstanz vorhanden ist, wird sie mit hoher Wahrscheinlichkeit einer bestimmten Klas-

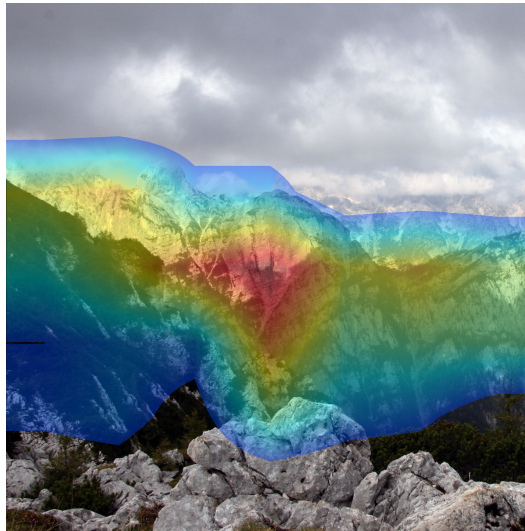


Abbildung 12.3: Grad-CAM-Erklärung für die Klassifikation eines Fotos als ‚Berge‘ durch das Modell ResNet-50 (76,74%). Von rot hin zu blau eingefärbten Bildbereichen nimmt der Einfluss auf die Klassifikation ab. Bildbereiche mit keinem oder geringem Einfluss auf die Klassifikation sind aus Gründen der Übersichtlichkeit nicht eingefärbt. Bildquelle: Privat / eigene Darstellung unter Verwendung von Grad-CAM (Selvaraju et al., 2017). Siehe Teil IV für Details.

se zugewiesen. Damit unterscheidet sich dieser Ansatz von den oben genannten Methoden: Anchors-Erklärungen typisieren die Extension jener Instanzen, welche die verankerten Merkmale teilen, und ermöglichen es dadurch, neue Instanzen zu beurteilen. Weiterhin kann die Methode mehrere Anker erzeugen, um das globale Verhalten des Klassifikationssystems abzudecken (Ribeiro et al., 2018, S. 1533). Das macht die Methode zu einem Kandidaten für den Erwerb von Verstehen_{Klasse}. Allerdings müsste die Interpretin dafür die Informationen mehrerer Anker synthetisieren, um ein ‚globales‘ Verständnis der für die fragliche Klassifikation relevanten Merkmale zu erlangen (Mann, 2025, S. 43).

Eine weitere Einschränkung, die auch Anchors betrifft, hat mit der Vagheit visueller Repräsentationen bzw. mit dem Problem der Extraktion zu tun (siehe Abschnitt 7.3.2). Trotz der Interpretationshilfen, die Salienzkarten bereitstellen, ist nicht immer eindeutig, welche Merkmale exemplifiziert werden. Denn Salienzkarten heben streng genommen lediglich Bild**bereiche** hervor, keine Bild**merkmale** (Achtibat et al., 2023; Ghassemi et al., 2021, S. 746; Rabold et al., 2020a,b; Rudin, 2019). Diese Bereiche können sowohl inerte als auch exemplifizierte Merkmale instanziierten. Akzentuiert die Salienzkarte aus Abbildung 12.3 die Farbe, Textur, Form und/oder Größe von dem, was innerhalb des hervorgehobenen Bildbereichs liegt? Salienzkarten allein können aufgrund der expressiven Einschränkungen visueller Repräsentationen darüber nichts aussagen. Das lässt sich anhand von Anchors

verdeutlichen: Bei der Erklärung von Bilddaten bestehen die Anker in Superpixeln der erklärten Instanz, d. h. in einem Bildausschnitt. Oft ist nicht eindeutig, wie diese Superpixel auf neue Instanzen übertragen werden können, da nicht klar ist, wie exakt die neue Instanz mit dem Anker übereinstimmen muss (Molnar, 2022, S. 214). Damit bewegen sich auch die visuellen Erklärungen der zweiten Gruppe in Richtung von SAMPLES (Mann, 2025, S. 43).

Zusammengefasst: Beispiele, die mithilfe von Attributionsmethoden generiert werden, haben den Vorteil, dass sie mehr Interpretationshilfen bereitstellen als prototypenbasierte Methoden. Das hilft dabei, inerte von exemplifizierte Merkmale zu unterscheiden. Andererseits reichen diese Interpretationshilfen nicht immer aus, um eine korrekte Interpretation zu gewährleisten. Nichtsdestotrotz scheinen lokale Methoden geeignet, um einzelne Klassifikationen verständlich zu machen. Sie sind hingegen nicht darauf ausgelegt, Verstehen_{Klasse} zu erzeugen (ebd.). Das ist bei den Methoden, die ich im nächsten Abschnitt betrachte, anders.

12.3.3 Gruppe 3: Ansätze, die ein künstliches Bild generieren

Um Verstehen_{Klasse} zu erwerben, braucht man ein Beispiel, das jene Merkmale exemplifiziert, die das Klassifikationssystem typischerweise nutzt, um Bilder der fraglichen Klasse zuzuweisen. Die zuvor besprochenen Methoden können einem Stakeholder prinzipiell epistemischen Zugang zu einigen dieser Merkmale verschaffen. Strebt man jedoch Verstehen_{Klasse} an, ist es naheliegend, direkt die Repräsentation einer Bildklasse zu visualisieren, die das Klassifikationssystem während des Trainings erlernt hat (ebd., S. 44).

Dafür kann man auf die in Abschnitt 3.2.2 beschriebene Feature-Visualisierung zurückgreifen. Vereinfacht ausgedrückt, erzeugt dieser Ansatz eines oder mehrere künstliche Bilder, durch die ein bestimmter Teil eines neuronalen Netzwerks maximal aktiviert wird. In gewisser Hinsicht kombiniert diese Methode die Vorteile von prototypenbasierten Ansätzen und Salienzkarten: Einerseits werden durch die Visualisierung der Repräsentation kontingente Merkmale, die Prototypen instanziierten, ausgesiebt. Andererseits wird dadurch ein Beispiel erzeugt, das Merkmale einer größeren Gruppe von Instanzen exemplifiziert, als es bei der Nutzung lokaler Methoden der Fall ist (ebd.).

Feature-Visualisierung kann prinzipiell auf jeden Teil eines neuronalen Netzes angewandt werden, von einzelnen Neuronen bis hin zur Ausgabeschicht, welche die finale Wahrscheinlichkeit einer Klasse ausgibt (Molnar, 2022, S. 244). Für Verstehen_{Klasse} bietet sich letzteres an (siehe Abbildung 12.4). Übertragen auf Exemplifikation heißt das Folgendes: Eine Feature-Visualisierung generiert Beispiele, die vornehmlich Merkmale instanziierten, die das System zur Klassifikation von Bildern einer Klasse heranzieht. Dies erleichtert beide Formen der Interpretation: Zum einen macht der Ansatz explizit, worin die typisierte Extension besteht (nämlich in typischen Instanzen der fraglichen Bildklasse). Zum anderen instanziiert die resultierende Visualisierung weniger inerte Merkmale, da prinzipiell nur solche Eigenschaften visualisiert werden, die sich positiv auf die fragliche Klassifikation auswirken. Dies hilft wiederum bei der Identifikation exemplifizierter Eigenschaften (Mann,

2025, S. 44).



Abbildung 12.4: Feature-Visualisierungen der Output-Neuronen des CaffeNet DNN für die Bildklassen ‚Maske‘, ‚Bibliothek‘, ‚Blattkäfer‘, ‚Lippenstift‘ und ‚Leuchtturm‘, erstellt mithilfe der Methode von Nguyen et al. (2016). Bildquelle: *Evolving Artificial Intelligence Lab* (2020). Siehe Teil IV für Details.

Feature-Visualisierungen lassen sich unterschiedlich stark steuern (eine Übersicht bieten Olah et al., 2017). Nimmt man wenige Einschränkungen vor, erhält man oftmals surrealistische Visualisierungen, die mehrere, unvollständige und perspektivisch verzerrte Instanzen der dargestellten Klasse zeigen. Solche Darstellungen können eine korrekte Interpretation erschweren, da sie sich nicht ohne Weiteres auf natürliche Bilder übertragen lassen – die exemplifizierten Merkmale sind für Menschen nicht gut nachvollziehbar. Greift man stärker ein, erhält man mehr oder minder natürlich aussehende Bilder, die in der Regel eine einzelne, vollständige Instanz der Bildklasse abbilden (wie in Abbildung 12.4). Von solchen Visualisierungen lässt sich leichter auf reale Bilder schließen. Der Nachteil stärkerer Einschränkungen besteht darin, dass sie zur Integration von Merkmalen führen können, die lediglich mit der Bildklasse korrelieren (ebd.). Dann ist abermals uneindeutig, welche der instanziierten Merkmale exemplifiziert und welche inert sind (Mann, 2025, S. 44). Aufgrund der Diskrepanz zwischen den von Menschen bzw. KI-Systemen verwendeten Klassifikationsstrategien bergen Beispiele, die inerte, aber zugleich natürlich wirkende Eigenschaften instanziierten, ein besonders hohes Risiko der Fehlinterpretation.

Trotz dieser Schwierigkeiten eignet sich Feature-Visualisierung besser als die zuvor diskutierten Methoden, um Verstehen_{Klasse} hervorzuheben. Die dadurch generierten Beispiele sind leichter zu interpretieren, weil sie weniger inerte Merkmale instanziierten. Außerdem zielen sie darauf ab, der Betrachterin epistemischen Zugang zu Merkmalen zu bieten, die typisch für eine ganze Bildklasse sind. Wird Verstehen_{Klasse} angestrebt, kommen diese Beispiele bereits nahe an EXEMPLARE heran (ebd.).

12.4 Einwände

In diesem Abschnitt weise ich drei mögliche Einwände gegen meinen Vorschlag zurück. Ich hoffe, dass selbst diejenigen Leser:innen, die von meinen Antworten auf diese Einwände nicht überzeugt sind, Folgendes zugestehen: Elgins Überlegungen zur Exemplifikation

lassen sich sinnvoll auf Erklärungen anwenden, die von XAI-Ansätzen erzeugt werden. Wenn man diese Erklärungen in Hinblick darauf analysiert, welche Merkmale sie exemplifizieren und welche Interpretationshilfen sie bereitstellen, kann man ihre Eignung für den Verstehensserwerb beurteilen (ebd., S. 47).

Alle drei Einwände betreffen die praktische Umsetzbarkeit von EXEMPLAREN: Demzufolge könnten EXEMPLARE nicht erzeugt werden, weil sich KI-Systeme auf unverständliche (Abschnitt 12.4.1) bzw. abstrakte (Abschnitt 12.4.2) Merkmale stützen oder weil das Generieren von EXEMPLAREN bereits das Verstehen voraussetzt, das sie hervorbringen sollen (Abschnitt 12.4.3). Betrachten wir diese Einwände im Einzelnen.

12.4.1 Einwand 1: Bildklassifikationssysteme stützen sich auf fremdartige Merkmale

Der erste Einwand lautet, dass EXEMPLARE wenigstens im Kontext der Bildklassifikation nicht generiert werden können, weil sich moderne Bildklassifikationssysteme auf Merkmale von Bildern stützen, die zu fremdartig sind, um für Menschen nachvollziehbar zu sein. Feindliche Beispiele untermauern diese Befürchtung (siehe Abschnitt 2.3.1). Zur Erinnerung: Im Kontext der Bildklassifikation handelt es sich bei feindlichen Beispielen klassischerweise um gezielt manipulierte Bilder, deren Klassifikation von der offensichtlich korrekten Bildklasse abweicht. Die Existenz feindlicher Beispiele scheint zu zeigen, dass Bildklassifikationssysteme (und, *mutatis mutandis*, neuronale Netze im Allgemeinen) Merkmale heranziehen, die zu fremdartig sind, um für Menschen verständlich zu sein (ebd.). Demzufolge wäre architektonische Opazität durch fremdartige Repräsentationen (Abschnitt 2.3.1) unüberwindbar.

Allerdings deuten manche Forschungsergebnisse darauf hin, dass die Umstände beeinflussen, wie Menschen feindliche Beispiele wahrnehmen. Zum Beispiel täuschten einige davon nicht nur KI-Systeme, sondern auch unter Zeitdruck stehende Studienteilnehmer:innen (Elsayed et al., 2019). In einer Studie von Zhenglong Zhou und Chaz Firestone konnten menschliche Proband:innen die Klassifikation feindlicher Beispiele korrekt vorhersagen, wenn sie aus einer begrenzten Anzahl von Bildklassen wählen mussten (Z. Zhou & Firestone, 2019). Darüber hinaus gibt es Hinweise, dass die Repräsentationen neuronaler Netze zumindest manchmal menschlichen Begriffen entsprechen (siehe Abschnitt 2.3.1). Angesichts dessen halte ich es für verfrüht, sich darauf festzulegen, dass moderne Bildklassifikationssysteme hoffnungslos opak sind. Wenn sich herausstellt, dass diese Systeme für Menschen nachvollziehbare Merkmale erlernen, können diese Merkmale durch EXEMPLARE exemplifiziert werden. Falls dem nicht so sein sollte, wären nicht nur EXEMPLARE unerreichbar, sondern wahrscheinlich auch jede andere für Menschen verständliche Erklärung von Bildklassifikationssystemen (oder anderen neuronalen Netzen). Der erste Einwand stellte dann kein für EXEMPLARE spezifisches Problem dar (Mann, 2025, S. 47).

12.4.2 Einwand 2: Bildklassifikationssysteme stützen sich auf abstrakte Merkmale

Der zweite Einwand besagt, dass EXEMPLARE unerreichbar bleiben, weil sich aktuelle Bildklassifikationssysteme auf abstrakte Merkmale von Bildern stützen, die nicht exemplifiziert werden können. In Bezug auf CNN argumentiert Cameron Buckner, dass diese Systeme eine abstrakte Repräsentation jeder Bildklasse erlernen (Buckner, 2018). Diese abstrakte Repräsentation ermögliche es, so genannte ‚Störfaktoren‘ (engl. *nuisance factors*), die sich in natürlichen Bildern befinden, auszuklammern. Störfaktoren definiert er als „repeatable and systematic sources of variation that are not diagnostic of decision success“ (Buckner, 2019, S. 9). Bei Bilddaten können solche Störfaktoren z. B. in Variationen von Größe, Position oder Rotation bestehen (ebd.). Dies legt nahe, dass nicht (allein) die riesige Menge an Klassifikationskriterien das Verstehen dieser Systeme verhindert. Vielmehr besteht die Schwierigkeit vornehmlich in deren Abstraktheit. Anders ausgedrückt: Das Problem ist nicht, dass die Repräsentation der Bildklasse ‚Rotkehlchen‘ eine überwältigende Anzahl von Rotkehlchengrößen, Körperhaltungen, Perspektiven usw. umfasst. Das Problem besteht eher darin, dass diese Repräsentation von Größe, Haltung oder Perspektive abstrahiert. Aufgrund der expressiven Einschränkungen visueller Repräsentationen scheint es unmöglich, solche abstrakten Eigenschaften durch ein visuelles EXEMPLAR zu exemplifizieren (siehe Abschnitt 7.3.2). Wie würde ein Beispiel der Klasse ‚Rotkehlchen‘ aussehen, das einen Vogel ohne bestimmte Größe oder räumliche Position zeigt (Mann, 2025, S. 47–48)?¹³⁹

Für die Zurückweisung dieses Einwands ist es hilfreich, zwischen der Funktionsweise neuronaler Netze und den Verstehensmöglichkeiten zu unterscheiden, die zum Verständnis dieser Systeme beitragen. Buckner (2018) befasst sich mit Ersterem. In diesem Zusammenhang ist der obige Einwand berechtigt: Abstrakte Bildmerkmale können nicht isoliert instanziiert werden, sondern nur in Kombination mit den oben genannten Störfaktoren. Jeder Versuch, solche Merkmale zu visualisieren, wird daher Falschheiten mit sich bringen (Mann, 2025, S. 48).

In diesem Kapitel geht es jedoch um Strategien, um KI-Systeme verständlich zu machen. Und wir haben bereits festgestellt, dass Falschheiten in dieser Hinsicht auch gelungen sein können. Zugegeben, abstrakte Merkmale allein lassen sich nicht instanziierten. Glücklicherweise können sie aber trotzdem exemplifiziert werden. Denn Störfaktoren können wie jedes andere inerte Merkmal, das ein Beispiel instanziiert, durch geeignete Interpretationshilfen heruntergespielt werden. Beispielsweise erfordert die Exemplifikation des abstrakten Merk-

¹³⁹Das ist natürlich eine Variante des bekannten Streits zwischen John Locke und George Berkeley. Sowohl Elgin (2017, Kap. 9) als auch Buckner (2018) diskutieren Lockes Passage über ‚die allgemeine Vorstellung eines Dreiecks‘ aus seinem *Essay Concerning Human Understanding* sowie George Berkeleys Antwort darauf im *Treatise Concerning the Principles of Human Knowledge*. Locke schrieb: „[...] the general idea of a triangle [...] must be neither oblique nor rectangle, neither equilateral, equicrural, nor scalenon; but all and none of these at once“ (Locke, 1979, Buch IV, Kap. 7, §9). Berkeley wiederum verspottete Lockes Beschreibung eines allgemeinen, aber inkonsistenten Dreiecks, von dem sich wohl niemand eine Vorstellung machen könne (Berkeley, 1949, §13–16 der Einleitung).

mals ‚einen Schnabel haben‘ (bzw. entsprechende Pixel besitzen) auch die Instanziierung einer bestimmten Schnabelposition (vgl. Elgin, 2017, S. 194–195). Solche inerten Merkmale können dann durch entsprechende Interpretationshilfen heruntergespielt werden. Das ist nichts Ungewöhnliches – auch eine hochdekorierte Ornithologin wird (und kann) keine abstrakte Rotkehlchenrepräsentation zur Erläuterung der charakteristischen Merkmale dieser Art heranziehen. Ich denke daher, dass prinzipiell nichts dagegen spricht, abstrakte Merkmale einer Bildklasse durch visuelle EXEMPLARE zu exemplifizieren. Im Gegenteil dürfte es oft schwierig sein, solche Merkmale zu verstehen, ohne dass sie in Form eines Beispiels konkretisiert werden (Mann, 2025, S. 48).

12.4.3 Einwand 3: Exemplifikation setzt Verstehen voraus

Der dritte Einwand lautet, dass das Erzeugen oder Auswählen eines EXEMPLARS bereits das Verstehen voraussetzt, das es hervorbringen soll. Eine Ornithologin kann nur deshalb auf bestimmte Merkmale von Rotkehlchen hinweisen (und andere herunterspielen), weil sie ein vorheriges Verständnis von charakteristischen Rotkehlchenmerkmalen hat. Das Gleiche scheint für EXEMPLARE zu gelten: Um sie erzeugen oder auswählen zu können, muss man die kontextuell relevanten Merkmale bereits kennen. Anders kann man nicht entscheiden, was durch ein EXEMPLAR exemplifiziert werden soll und was nicht (ebd.).

Um diesen Einwand zurückzuweisen, muss man zwischen zwei Arten von Beispielen differenzieren: Erstens gibt es Beispiele, die didaktischen Zwecken dienen und Verstehen vermitteln sollen, das eine andere Person (z. B. eine Expertin) bereits besitzt. Zweitens können Beispiele auch als Quelle für ein Verständnis dienen, das vorher nicht vorhanden war (Elgin, 2017, S. 190).

Im zweiten Fall werden Erklärbarkeitsmethoden oder andere Verstehensermöglicher von Expert:innen verwendet, um ein besseres Verständnis von KI-Systemen oder einigen ihrer Aspekte zu erlangen. Eine Serie von Aufsätzen aus der Publikation *Distill* ist dafür ein gutes Beispiel (Cammarata et al., 2020a). Die Autor:innen dieser Aufsätze analysierten tausende Neuronen eines CNN, um detailliert zu verstehen, wie das System Bilddaten verarbeitet. Dafür griffen die Wissenschaftler:innen unter anderem auf die in Abschnitt 12.3.3 beschriebene Feature-Visualisierung zurück und visualisierten für jedes Neuron jene Merkmale, auf die es am stärksten reagiert (Olah et al., 2017). Mit anderen Worten: Sie generierten ein Beispiel, das ihnen ein völlig neues Verständnis der Repräsentationen des Netzwerks ermöglichte. Das funktioniert natürlich nur, wenn das Beispiel seinen Referenten adäquat repräsentiert. Dafür muss es beispielsweise aus geeigneten Methoden hervorgehen und ggf. mit bestehendem Vorwissen abgeglichen werden (vgl. Elgin, 2017, S. 190). Das ist wiederum keine Besonderheit von EXEMPLAREN, sondern gilt gleichermaßen für andere in XAI-Kontexten verwendete Verstehensermöglicher.

Sofern die eingesetzten XAI-Ansätze zuverlässig sind, können also EXEMPLARE generiert oder ausgewählt werden, ohne das Verstehen vorauszusetzen, das sie herbeiführen sollen. Wenn dann das erforderliche Verstehen vorhanden ist, können EXEMPLARE auch

verwendet werden, um dieses Verständnis anderen zu vermitteln – zugeschnitten auf die kontextuellen Anforderungen des jeweiligen Stakeholders. EXEMPLARE kommen damit auch als didaktische Verstehensermöglicher für Nichtfachleute in Frage und können auch der zuerst genannten Rolle von Beispielen gerecht werden (Mann, 2025, S. 48).

Dieser Vorschlag zu wünschenswerten Eigenschaften von Verstehensermöglichern beendet meine Untersuchung von Verstehen in XAI-Kontexten. Mit meinem Begriff instrumentellen Verstehens habe ich ein wichtiges Ziel von erklärbarer KI präzisiert. EXEMPLARE stellen ein besonders vorteilhaftes Mittel dar, um dieses Ziel zu erreichen.

13 Konklusion

In Kapitel 1 habe ich das Black-Box-Problem als Gefährdung der Erfüllung gesellschaftlicher Desiderata in KI-Kontexten charakterisiert, die im unzureichenden Verständnis der involvierten KI-Systeme begründet ist. Dieses Black-Box-Problem, und die Lösungsversuche des informatischen Forschungsfelds der erklärbaren Künstlichen Intelligenz (XAI), motivierten die beiden Forschungsfragen, denen ich in dieser Arbeit nachgegangen bin. Die erste lautete:

- (a) Welche Form des Verstehens ermöglicht Stakeholdern die Erfüllung gesellschaftlicher Desiderata in XAI-Kontexten?

Um diese Frage zu beantworten, habe ich einen an Rudolf Carnaps Methode der Explikation angelehnten, begriffsgestaltenden Ansatz gewählt. Mein Explikandum war der Alltagsbegriff von Verstehen, wie er in der XAI-Debatte üblicherweise verwendet wird (,Verstehen *simpliciter*'). Mein Explikat habe ich aufgrund seines zweckgebundenen Charakters als ,instrumentelles Verstehen' bezeichnet. Das Ergebnis meiner Untersuchung lautete, knapp ausgedrückt, dass instrumentelles Verstehen genau dann gegeben ist, wenn ein Stakeholder ein gerechtfertigtes, moderat faktives Überzeugungssystem über systembezogene Aspekte besitzt sowie über verstehensbezogene Fähigkeiten verfügt. Drei Kontextfaktoren wirken sich auf diese Bedingungen aus: Zum einen beeinflussen die *kontextuellen Risiken und Chancen* den erforderlichen Grad der Rechtfertigung. Zum anderen entscheidet das *Desideratum* des Stakeholders, ggf. ergänzt um seine *Rolle* als limitierenden Faktor, welche Überzeugungen des Überzeugungssystems wahr oder annähernd wahr sein müssen und über welche Fähigkeiten der Stakeholder verfügen muss. Wird den Bedingungen instrumentellen Verstehens in einem konkreten Kontext entsprochen, so mein Vorschlag, ermöglicht dies dem instrumentell verstehenden Stakeholder, sein Desideratum epistemisch (und in manchen Fällen auch substanziell) zu erfüllen.

Um die begriffliche Kontinuität zwischen Explikandum und Explikat sicherzustellen, habe ich mich mit zentralen erkenntnis- und wissenschaftstheoretischen Debatten zu Verstehen auseinandergesetzt. Im Zuge dieser Diskussion habe ich nicht nur instrumentelles Verstehen charakterisiert, sondern auch diverse Eigenschaften von Verstehen *simpliciter* herausgearbeitet, die in der gegenwärtigen philosophischen Literatur kontrovers diskutiert werden:

- Entgegen einer verbreiteten Auffassung sind nicht nur objektuales Verstehen, sondern auch propositionales Verstehen (Verstehen-dass) und interrogatives Verstehen

(Verstehen-w) holistisch und betreffen stets Objekte einer gewissen Komplexität, niemals eine einzelne, simple Proposition. Der Grund dafür ist, dass das Begreifen von Zusammenhängen in einer hinreichend komplexen Informationsmenge ein notwendiges Merkmal von Verstehen ist. Entsprechend ist auch propositionales Verstehen, anders als üblicherweise angenommen wird, gradierbar.

- Anders als diverse Autor:innen vorgeschlagen haben, ist es plausibel, dass Verstehen Überzeugung erfordert. Diese Position ist kompatibel damit, dass Wissenschaftler:innen mitunter Theorien verstehen, die sie nicht für wahr halten. Sie ist auch damit vereinbar, dass Wissenschaftler:innen Aspekte der Welt mithilfe einer idealisierten oder abstrahierten Theorie verstehen, die streng genommen falsch ist. Der Schlüssel für die Zurückweisung vorgeblicher Gegenbeispiele bestand in der Unterscheidung zwischen faktischem und symbolischem Verstehen, also dem Verstehen eines Aspekts der Welt versus dem Verstehen einer Repräsentation. Weiterhin steht eine notwendige Überzeugungsbedingung nicht im Widerspruch damit, dass nicht-propositionale Repräsentationen verstanden werden können: Manche nicht-propositionalen Repräsentationen haben propositionale Inhalte, und auch nicht-propositionale Inhalte können in Propositionen übersetzt oder in propositionale Strukturen integriert werden.
- Anstelle dieses gut bewältigbaren Problems der Propositionalität habe ich eine andere, für XAI-Kontexte viel wichtigere Schwierigkeit herausgearbeitet: das Problem der Extraktion. Dieses Problem ergibt sich daraus, dass die Inhalte visueller Repräsentationen oftmals schwierig zu bestimmen sind, weil Abbildungen inhärente expressive Limitationen aufweisen und zunächst korrekt interpretiert werden müssen. Diese Vagheit erschwert z. B. den Erwerb wahrer Überzeugungen auf Grundlage von Salienzarten, die in XAI-Kontexten häufig zum Einsatz kommen.
- Verstehen ist moderat faktiv, es ist also sowohl mit peripheren falschen Überzeugungen kompatibel als auch mit zentralen, lediglich approximativ wahren Überzeugungen. Anders als viele Autor:innen annehmen, lässt sich dies jedoch nicht durch den Verweis auf die Tatsache zeigen, dass Wissenschaftler:innen häufig Theorien verwenden, die streng genommen falsch sind: Die Verwendung falscher Theorien ist mit einem strengen Faktivismus vereinbar. Gegen eine streng faktive Position spricht vielmehr die alltägliche Praxis der Zuschreibung von Verstehen und die Tatsache, dass eine solche Position praktisch unerreichbare Standards für Verstehen setzen würde.
- Verstehen erfordert unter anderem das Verfügen über verstehensbezogene Fähigkeiten. Ich habe konkurrierende Positionen zurückgewiesen, die Fähigkeiten als nicht notwendig bzw. als hinreichend für Verstehen auffassen. Zumindest basale private Fähigkeiten, wie das Erkennen der Bedeutung von Symbolen, sind plausiblerweise

notwendig für Verstehen. Allerdings sind selbst umfangreiche Fähigkeiten dafür nicht hinreichend – das Verstehenssubjekt braucht auch ein moderat faktives, gerechtfertigtes Überzeugungssystem. Der Pool an Fähigkeiten, die charakteristisch für Verstehen sein können, ist umfangreich. Im Rahmen einer Literaturschau habe ich eine Vielzahl verstehensbezogener Fähigkeiten identifiziert und anhand von sechs Clustern systematisiert: erkennen, bewerten, erklären, schlussfolgern, intervenieren und entwickeln.

- Schließlich erfordert Verstehen, wie auch Wissen, Rechtfertigung. Vorgebliche Fälle von Verstehen ohne Rechtfertigung habe ich zurückgewiesen, indem ich die Begreifens- und Rechtfertigungskomponente von Verstehen klar voneinander getrennt und faktisches bzw. symbolisches Verstehen strikt auseinandergehalten habe. Weil Verstehen das Begreifen von Zusammenhängen erfordert, ist dafür eine internalistische, hybride Theorie der Rechtfertigung besonders attraktiv, die fundamentalistische und kohärentistische Elemente kombiniert.

Viele Eigenschaften von Verstehen *simpliciter* treffen gleichermaßen auf instrumentelles Verstehen zu: Beide haben ausschließlich komplexe Objekte, erfordern Rechtfertigung, Überzeugung sowie das Verfügen über Fähigkeiten und sind moderat faktiv. In anderen Hinsichten weist instrumentelles Verstehen allerdings Besonderheiten auf:

- Instrumentelles Verstehen ist im Gegensatz zu Verstehen *simpliciter* nicht gradierbar. Es entspricht vielmehr einem kontextspezifischen Verstehensgrad, der die Erfüllung eines bestimmten Desideratums ermöglicht. Wird dieser Grad nicht erreicht, verfehlt das Verstehen seinen Zweck und ist nicht instrumentell. Wird er überschritten, verbessert dies zwar den Verstehensgrad *simpliciter*, ist aus instrumenteller Perspektive aber überflüssig.
- Bei instrumentellem Verstehen entscheidet der Kontext, welche Überzeugungen zentral sind und daher wahr oder wenigstens näherungsweise wahr sein müssen, um das Kriterium der moderaten Faktizität zu erfüllen. Während es bei Verstehen *simpliciter* in einem relevanten Sinne kontextunabhängige zentrale bzw. periphere Überzeugungen zu geben scheint, macht der zweckgebundene Charakter instrumentellen Verstehens eine kontextsensitive Bestimmung der Zentralität von Verstehensüberzeugungen erforderlich.
- Instrumentelles Verstehen kann verstehensbezogene Fähigkeiten erfordern, die für Verstehen *simpliciter* nicht notwendig zu sein scheinen. Auch diese Abweichung ist durch den theoretischen Zweck des Begriffs motiviert: Mitunter erfordert die durch Verstehen medierte Erfüllung eines Desideratums Fähigkeiten, die zu anspruchsvoll sind, um als notwendig für Verstehen *simpliciter* gelten zu können.

Nachdem der epistemische Zielzustand in den hier relevanten XAI-Kontexten näher bestimmt war, stellte sich die Frage, wie dieser Zustand erreicht werden kann. Ich habe vorgeschlagen, neben informatischen XAI-Erklärungen sowie wissenschaftlichen oder alltäglichen Erklärungen im philosophischen Sinne auch andere Vehikel für den Verstehenswerb in den Blick zu nehmen. Dazu zählen so unterschiedliche Dinge wie diagrammatische Zusammenfassungen relevanter Eigenschaften von Trainingsdaten, Datenblätter mit technischen Spezifikationen, die Erklärung einer Expertin oder auch offizielle Regelwerke und Gesetzestexte. Kurz: Instrumentelles Verstehen kann prinzipiell mithilfe verschiedenster Repräsentationen erworben werden, so lange diese Auskunft über relevante systembezogene Aspekte geben. Ich habe diese vielfältigen Repräsentationen unter dem Begriff des Verstehensermöglicher zusammengefasst. Doch was macht einen Verstehensermöglicher geeignet, um instrumentelles Verstehen hervorzubringen? Dies führte mich zu meiner zweiten Forschungsfrage:

- (b) Welche Eigenschaften muss ein Verstehensermöglicher aufweisen, um der Erlangung instrumentellen Verstehens zuträglich zu sein?

Ich habe vorgeschlagen, dass effektive Verstehensermöglicher einem Stakeholder das Verständnis von für instrumentelles Verstehen relevanten systembezogenen Aspekten ermöglichen, indem sie diese Aspekte – im Sinne von Catherine Elgins zugehöriger Theorie – exemplifizieren. Ich habe vorteilhafte Verstehensermöglicher als EXEMPLARE bezeichnet und durch fünf Eigenschaften charakterisiert, die eine korrekte Interpretation begünstigen: EXEMPLARE (1) heben exemplifizierte Merkmale hervor, (2) spielen inerte Merkmale herunter, (3) weisen einen klaren Geltungsbereich auf und exemplifizieren (4) kontextuell relevante sowie (5) für Menschen nachvollziehbare Merkmale. Um die praktische Anwendbarkeit meines Vorschlags zu illustrieren, habe ich diese Überlegungen auf diverse visuelle Erklärungen angewandt, die verwendet werden, um die Ausgaben von Bildklassifikationssystemen verständlich zu machen. Mein Ergebnis war, dass die meisten Visualisierungen zumindest in Bezug auf ein exemplarisches epistemisches Ziel – nämlich das Verstehen der Klassifikationskriterien für eine gesamte Bildklasse – die Kriterien von EXEMPLAREN nicht erfüllen.

Zusammengefasst sind die beiden wichtigsten Ergebnisse meiner Untersuchung die folgenden: Erstens habe ich auf Grundlage aktueller erkenntnis- und wissenschaftstheoretischer Debatten einen Verstehensbegriff gestaltet, welcher der häufig behaupteten Mediatorrolle von Verstehen in XAI-Kontexten gerecht wird: *instrumentelles Verstehen*. Mithilfe meiner Definition lässt sich grundsätzlich sowie in konkreten Kontexten erklären, wie eine bestimmte Form des Verstehens die Erfüllung diverser Desiderata ermöglichen kann. Mein Vorschlag konkretisiert dadurch die allgemeinen Verweise auf den Alltagsbegriff, die in der Erklärbarkeitsliteratur üblicherweise erfolgen, und präzisiert ein zentrales epistemisches Ziel in XAI-Kontexten. Wenngleich sich mein Vorschlag vornehmlich an ein philosophisches Fachpublikum richtet, bietet mein Begriffsvorschlag auch Anknüpfungspunkte für die empirische XAI-Forschung. Insbesondere verstehensbezogene Fähigkeiten

sind attraktive Kandidaten für die Operationalisierung instrumentellen Verstehens und damit für dessen empirische Evaluation im Kontext erklärbarer KI.

Zweitens habe ich Eigenschaften von epistemisch besonders vorteilhaften Verstehensermöglichern herausgearbeitet, die den Erwerb instrumentellen Verstehens begünstigen. Diese Eigenschaften stellen einen theoretisch fundierten Ausgangspunkt dar, um bestehende XAI-Erklärungen zu evaluieren und zu verbessern. Sie können auch dabei helfen, neue Erklärbarkeitsmethoden zu entwickeln, deren Ausgaben sich EXEMPLAREN annähern. Weiterhin stellen EXEMPLARE eine Antwort auf das Problem der Vagheit visueller XAI-Erklärungen dar, indem sie die korrekte Interpretation solcher Repräsentationen begünstigen. Ich hoffe dadurch auch, XAI-Forscher:innen für diese Problematik zu sensibilisieren und dazu anzuregen, höhere Ansprüche an die Interpretierbarkeit von Salienzkarten und anderen in XAI verwendeten visuellen Repräsentationen zu stellen.

Nichtsdestotrotz ist die Untersuchung von Verstehen in XAI-Kontexten damit keineswegs abgeschlossen. Vielmehr haben sich im Verlauf dieser Arbeit eine Reihe weiterführender Forschungsfragen ergeben. Eine davon betrifft die in Verstehen involvierte epistemische Einstellung. Ich habe argumentiert, dass eine Überzeugungsbedingung für (instrumentelles) Verstehen gut mit vorgeblichen Fällen von Verstehen ohne Überzeugung umgehen kann. Doch stützte sich meine Argumentation in Teilen auch darauf, dass die stattdessen häufig vorgeschlagene Einstellung der Akzeptanz in manchen Hinsichten nicht ausgereift genug ist, um eine attraktive Alternative zu Überzeugung darzustellen. Konkret ist noch unklar, wie nicht-propositionale Akzeptanz aussehen könnte. Es wäre spannend, genauer zu untersuchen, ob Verstehen mitunter nicht-propositionale Einstellungen erfordert (wie sie zum Beispiel in visuellem Denken involviert sein könnten). Wenn dem so sein sollte, wäre eine Überzeugungsbedingung womöglich nicht notwendig für instrumentelles Verstehen bzw. Verstehen *simpliciter*.

Ein verwandter Aspekt betrifft die phänomenale und affektive Komponente von Verstehen bzw. Begreifen, die in den letzten Jahren vermehrt diskutiert worden ist. Ich bin zuversichtlich, dass diese Komponenten im Rahmen der Fähigkeitsbedingung durch meinen Begriff instrumentellen Verstehens abgedeckt werden. Allerdings verdient dieser spannende Aspekt eine ausführlichere Betrachtung, als ich in dieser Arbeit leisten konnte.

Auch die Kompatibilität von (instrumentellem) Verstehen mit lediglich näherungsweise wahren Überzeugungen könnte weiter exploriert werden. Wie sollte die Approximation an die Wahrheit im Kontext von Verstehen bestenfalls aufgefasst werden? Wie stark müssen sich zentrale Überzeugungen der Wahrheit annähern, um ausreichend faktiv zu sein? Um das subtile Verhältnis von Verstehen zu den Fakten weiter aufzuklären, wäre eine dezidierte Auseinandersetzung mit Debatten zur Wahrheitsnähe hilfreich.

Weiterhin habe ich ein schwieriges Thema, den Kontext, auf drei für mich besonders relevante Kontextfaktoren (das Desideratum und die Rolle des Stakeholders sowie die kontextuellen Risiken bzw. Chancen) reduziert. Aufgrund der Vielfalt der gegenwärtigen und künftig zu erwartenden Anwendungsbereiche von KI könnte es sein, dass diese Betrachtung der Komplexität von XAI-Kontexten nicht gerecht wird. Eine genauere Analyse der Auswir-

kungen weiterer Kontextfaktoren auf die Bedingungen von (instrumentellem) Verstehen wäre nicht nur in Hinblick auf die Kontextsensitivität von Verstehen aufschlussreich, sondern auch wichtig, um weitere philosophisch fundierte Anforderungen an XAI-Erklärungen und andere Verstehensermöglicher formulieren zu können.

Es wäre außerdem spannend, meine Überlegungen mit Forschung aus der sozialen Erkenntnistheorie zu verknüpfen. Die epistemische Arbeit in XAI-Kontexten ist, wie in anderen komplexen Domänen auch, arbeitsteilig organisiert. Viele Stakeholder sind daher epistemisch abhängig voneinander. Wie in dieser Arbeit deutlich wurde, kann die daraus resultierende Opazität zum Problem werden. Gleichzeitig birgt epistemische Arbeitsteilung die wohl einzige Chance, hochgradig komplexe KI-Systeme verständlich genug zu machen, um relevante gesellschaftliche Desiderata zu erfüllen. Wir können uns ideale XAI-Kontexte als Ökosysteme vorstellen, in denen sich die Expertise verschiedener Stakeholder ergänzt, die unterschiedlichen relevanten Quellen von Opazität jeweils durch geeignete Verstehensermöglicher überwunden werden und dadurch die Erfüllung von Desiderata trotz individueller, durchaus legitimer Verstehenslücken gewährleistet wird – weil die Expertise der anderen diese Verstehenslücken absichert. Ich denke, dass solche sozialepistemologischen Überlegungen zentral für den Verstehenserwerb in XAI-Kontexten sind, und es wäre spannend, meinen Begriff instrumentellen Verstehens in Hinblick darauf weiterzuentwickeln.

Auch meine Analyse vorteilhafter Verstehensermöglicher bietet interessante Anknüpfungspunkte für weitere Forschung. Elgins Theorie der Exemplifikation kann auf alle erdenklichen Repräsentationen von für instrumentelles Verstehen relevanten Inhalten angewandt werden. Doch konnte ich dies aus Platzgründen nur kurz diskutieren. Es wäre spannend, meine Überlegungen zu EXEMPLAREN auf andere visuelle Repräsentationen (neben visuellen Erklärungen im Kontext der Bildklassifikation) bzw. nicht-visuelle Repräsentationen (z. B. Systemdokumentationen, linguistische Erklärungen) anzuwenden und zu prüfen, ob sich daraus weitere wünschenswerte Eigenschaften von EXEMPLAREN ableiten lassen.

Schließlich verdient eine in meinen Augen wichtige Art von Verstehensermöglicher eine ausführliche Betrachtung – wir können sie ‚dynamische Verstehensermöglicher‘ nennen. Idealerweise sind Verstehensermöglicher keine statischen Repräsentationen, sondern lassen iterative Nachfragen durch die Adressatin und dynamische Anpassungen an kontextuelle Veränderungen zu (z. B. an das Hintergrundwissen der Adressatin, das Vorhandensein von Zeitdruck oder kontextuelle Risiken). Zudem könnten Verstehensermöglicher die Rolle einer Lehrerin oder eines Tutors einnehmen und einen Stakeholder gezielt dabei unterstützen, das für ihn relevante Verständnis zu erwerben – sei es durch gezielte Nachfragen oder das Bereitstellen relevanter Hintergrundinformationen. Eine solche didaktische Perspektive auf Verstehensermöglicher bietet abermals spannende Anknüpfungspunkte an sozialepistemologische Überlegungen zu Verstehen, aber auch an tugenderkenntnistheoretische Überlegungen (etwa zu intellektuellen Tugenden des Lehrens).

Als ich die Arbeit an dieser Dissertation begann, war Künstliche Intelligenz längst ein etabliertes Forschungsfeld und auch gängiges Thema in Science-Fiction und Popkultur. Doch für die meisten Menschen blieb diese Technologie eher abstrakt und betraf, wenn überhaupt, die fernere Zukunft. Dieses Bild hat sich innerhalb kürzester Zeit stark gewandelt. Mittlerweile ist Künstliche Intelligenz keine unwirkliche Zukunftstechnologie mehr, sondern für viele ein alltägliches Werkzeug geworden. Beinahe täglich machen Meldungen über bizarres Verhalten großer Sprachmodelle, ‚KI-Psychosen‘ oder ‚KI-Slop‘ die Runde, regelmäßig werden neue Fälle des missbräuchlichen Einsatzes von KI durch Betrüger:innen oder gar Behörden bekannt, und in Talkshows und Universitäten wird hitzig über die Transformation der Arbeitswelt und des Bildungswesens durch KI-Tools und KI-Agenten diskutiert. Der Einfluss von Künstlicher Intelligenz – im Positiven wie Negativen – ist immens und es ist schwierig vorherzusagen, welche Veränderungen uns künftig erwarten. Angesichts dieser rasanten technischen und gesellschaftlichen Entwicklungen scheint es aus philosophischer Sicht besonders wichtig, einen Schritt zurückzutreten und grundsätzlich zu fragen, wie KI-Systeme gestaltet werden sollten, damit sie relevanten Anforderungen gerecht werden. Diese Arbeit hat gezeigt, dass die Reflexion über die epistemische Dimension solcher Desiderata dafür einen zentralen Ausgangspunkt darstellt.

Teil IV

Appendix

Abbildungsverzeichnis

1.1	Analyseschema für XAI-Kontexte	6
2.1	Bereiche der informatischen KI-Forschung	12
2.2	Entscheidungsbaum	14
2.3	Künstliches neuronales Netz und Perzeptron	15
2.4	Taxonomie von Opazitätsquellen	22
3.1	Taxonomie von XAI-Methoden	33
3.2	Grad-CAM-Erklärung ‚Schwein‘	35
3.3	LIME-Erklärung ‚Hängebrücke‘	36
5.1	Dimensionen der Gradierbarkeit von Verstehen	94
7.1	Grad-CAM-Erklärung ‚Bobtail‘	128
12.1	Beispiel der Bildklasse ‚Rotkehlchen‘	233
12.2	MMD-Critic-Erklärung: Prototypen und Kritiken für zwei Bildklassen . . .	250
12.3	Grad-CAM-Erklärung ‚Berge‘	252
12.4	Feature-Visualisierungen verschiedener Bildklassen	254

Tabellenverzeichnis

9.1	Fähigkeiten des Erkennen-Clusters	166
9.2	Fähigkeiten des Bewerten-Clusters	168
9.3	Fähigkeiten des Erklären-Clusters	169
9.4	Fähigkeiten des Schlussfolgern-Clusters	171
9.5	Fähigkeiten des Intervenieren-Clusters	173
9.6	Fähigkeiten des Entwickeln-Clusters	174

Ergänzende Angaben

Angaben zum Einsatz von KI-gestützten Hilfsmitteln

Teile der vorliegenden Dissertation (Abschnitt 2.3, Kapitel 9 sowie Kapitel 12) wurden in abgewandelter Form bereits als englischsprachige Aufsätze publiziert (Mann, 2025; Mann et al., 2023; Speith et al., 2024). Für die Integration dieser Arbeiten in die Dissertation habe ich mithilfe von DeepL jeweils eine Übersetzung ins Deutsche angefertigt. Diese Übersetzungen habe ich gründlich geprüft, korrigiert und für die Integration in die Arbeit überarbeitet. Ich habe keine längeren mit KI übersetzten Textabschnitte unverändert übernommen.

Außerdem habe ich mithilfe von ChatGPT Python-Code generiert bzw. adaptiert, mit dem ich die Bildklassifikationssysteme ResNet-50 und Inception V3 sowie die Erklärbarkeitsmethoden LIME und Grad-CAM lokal auf meinem Rechner implementiert habe. Mit diesem Code habe ich dann Fotos von den genannten Systemen klassifizieren lassen und mit den genannten Erklärbarkeitsmethoden vier visuelle Erklärungen für diese Klassifikationen generiert (Abbildungen 3.2, 3.3, 7.1 und 12.3).

Angaben zum Ursprung einzelner Textabschnitte

- Das dieser Arbeit vorangestellte Epigraph stammt aus einem Brief von George Sand (1884, S. 196).
- Teile dieser Arbeit wurden in abgewandelter Form bereits als wissenschaftliche Aufsätze publiziert:
 - **Abschnitt 2.3** basiert auf Mann et al. (2023). Copyright ©2023, IEEE. Der Text wurde für diese Dissertation ins Deutsche übersetzt sowie inhaltlich überarbeitet, gekürzt und erweitert. Zahlreiche Änderungen gegenüber dem Original wurden vorgenommen.
 - **Kapitel 9** basiert auf Speith et al. (2024). Der Aufsatz ist lizenziert via ‚CC BY 4.0 Namensnennung 4.0 International‘ (<https://creativecommons.org/licenses/by/4.0/>). Der Text wurde für diese Dissertation ins Deutsche übersetzt sowie inhaltlich überarbeitet, gekürzt und erweitert. Zahlreiche Änderungen gegenüber dem Original wurden vorgenommen.

- **Kapitel 12** basiert auf Mann (2025). Der Aufsatz ist lizenziert via ‚CC BY 4.0 Namensnennung 4.0 International‘ (<https://creativecommons.org/licenses/by/4.0/>). Der Text wurde für diese Dissertation ins Deutsche übersetzt sowie inhaltlich überarbeitet, gekürzt und erweitert. Zahlreiche Änderungen gegenüber dem Original wurden vorgenommen.

Angaben zu Abbildungen

- **Abbildung 1.1:** Es handelt sich um eine von mir erstellte Abbildung unter Verwendung der folgenden, laut Lizenzangabe der jeweiligen Datei als Public-Domain- bzw. via ‚CC0 1.0 Universell‘ (<https://creativecommons.org/publicdomain/zero/1.0/>) lizenzierten Vektorgrafiken aus SVG Repo (www.svgrepo.com):
 - ‚Person Profile User Account Group‘, hochgeladen von SVG Repo, URL: <https://www.svgrepo.com/svg/392285/person-profile-user-account-group-people>.
 - ‚Loudspeaker 6 SVG Vector‘, Autor: Icooon Mono, URL: <https://www.svgrepo.com/svg/479596/loudspeaker-6>.
 - ‚Computer Tv SVG Vector‘, hochgeladen von SVG Repo, URL: <https://www.svgrepo.com/svg/287480/computer-tv>.
 - ‚Light Bulb Idea SVG Vector‘, hochgeladen von SVG Repo, URL: <https://www.svgrepo.com/svg/262163/light-bulb-idea>.
- **Abbildung 2.1, Abbildung 2.2 und Abbildung 5.1:** Es handelt sich um von mir erstellte Abbildungen.
- **Abbildung 2.3a:** Es handelt sich um eine von mir erstellte Abbildung adaptiert von der Public-Domain-Abbildung ‚Backpropagation‘ des Autors ToXiC Dz der arabischen Wikipedia (via Wikimedia Commons), URL: <https://commons.wikimedia.org/wiki/File:Backpropagation.png>.
- **Abbildung 2.3b:** Es handelt sich um eine von mir erstellte Abbildung, adaptiert von einem Ausschnitt von Abbildung 2 in Maier et al. (2019, S. 88). Der Aufsatz ist lizenziert via ‚CC BY 4.0 Namensnennung 4.0 International‘ (<https://creativecommons.org/licenses/by/4.0/>).
- **Abbildung 2.4:** Es handelt sich um eine von mir erstellte Abbildung adaptiert von Abbildung 1 in Mann et al. (2023, S. 339). Copyright ©2023, IEEE.
- **Abbildung 3.1:** Es handelt sich um eine von mir erstellte Abbildung adaptiert von Abbildung 3 in Speith (2022, S. 2242). Der Aufsatz ist lizenziert via ‚CC BY 4.0 Namensnennung 4.0 International‘ (<https://creativecommons.org/licenses/by/4.0/>).

-
- **Abbildung 3.2a:** Es handelt sich um ein von mir aufgenommenes Foto.
 - **Abbildung 3.2b, Abbildung 7.1 und Abbildung 12.3** basieren auf einer für meine Zwecke angepassten Implementierung von Grad-CAM (Selvaraju et al., 2017) auf Basis der Implementierung von Gildenblat (2026) und unter Verwendung von selbst aufgenommenen Fotos. Grundlage bildete die Klassifikation dieser Fotos durch das Modell ResNet-50, das ich auf Basis der Implementierung von He (2017) implementiert habe. Ich habe für meine Untersuchung jeweils Anpassungen am Python-Code vorgenommen. Verwendung von Abbildung 7.1 mit freundlicher Genehmigung von Henrys Besitzerin sowie von der Technischen Universität Dortmund, Referat 1 Hochschulkommunikation.
 - **Abbildung 3.3a und Abbildung 3.3b** basieren auf einer für meine Zwecke angepassten Implementierung von LIME (Ribeiro et al., 2016) auf Basis der Implementierung von Ribeiro (2024) und unter Verwendung eines selbst aufgenommenen Fotos. Grundlage bildete die Klassifikation dieses Fotos durch das Modell Inception V3, das ich auf Basis der Implementierung von Tensorflow (2026) implementiert habe. Ich habe für meine Untersuchung jeweils Anpassungen am Python-Code vorgenommen.
 - **Abbildung 12.1:** Es handelt sich um die Abbildung ‚European Robin (Erithacus rubecula)‘ des Autors botaflo (via Wikimedia Commons), URL: https://commons.wikimedia.org/wiki/File:Erithacus_rubecula_108174470.jpg. Die Abbildung ist lizenziert via ‚CC0 1.0 Universell‘ (<https://creativecommons.org/publicdomain/zero/1.0/>).
 - **Abbildung 12.2:** Es handelt sich um Abbildung 2 aus B. Kim et al. (2016, S. 8). Verwendung der Abbildung mit freundlicher Zustimmung von Been Kim und Rajiv Khanna.
 - **Abbildung 12.4:** Es handelt sich um eine Abbildung des Evolving Artificial Intelligence Lab (2020), URL: <https://github.com/Evolving-AI-Lab/synthesizing/blob/master/examples/example1.jpg>. Zu sehen sind Feature-Visualisierungen der Output-Neuronen des CaffeNet DNN. Verwendung der Abbildung mit freundlicher Zustimmung von Anh Totti Nguyen.

Literatur

- Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Židek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis & John M. Jumper (2024). „Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3“. In: *Nature*, Bd. 630, Nr. 8016, S. 493–500. DOI: 10.1038/s41586-024-07487-w.
- Abu-Ghazaleh, Nael, Dmitry Ponomarev & Dmitry Evtvyushkin (2019). „How the Spectre and Meltdown Hacks Really Worked“. In: *IEEE Spectrum*, Bd. 56, Nr. 3, S. 42–49. DOI: 10.1109/mspec.2019.8651934.
- Achinstein, Peter (1983). *The Nature of Explanation*. New York: Oxford University Press.
- Achtibat, Reduan, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek & Sebastian Lapuschkin (2023). „From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation“. In: *Nature Machine Intelligence*, Bd. 5, Nr. 9, S. 1006–1019. DOI: 10.1038/s42256-023-00711-8.
- Alvarez, Maria (2010). *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford, New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780199550005.001.0001.
- Alvarez, Maria & Jonathan Way (2024). „Reasons for Action: Justification, Motivation, Explanation“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta & Uri Nodelman. Herbst-Edition. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/fall2024/entries/reasons-just-vs-expl/>.
- Angelov, Plamen P., Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold & Peter M. Atkinson (2021). „Explainable Artificial Intelligence: An Analytical Review“. In: *WIREs Data Mining and Knowledge Discovery*, Bd. 11, Nr. 5, Art. Nr. e1424, S. 1–13. DOI: 10.1002/widm.1424.
- Angwin, Julia, Jeff Larson, Surya Mattu & Lauren Kirchner (23. Mai 2016). *There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*. ProPublica. URL: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (besucht am 29. 09. 2025).

- Arbesman, Samuel (2017). *Overcomplicated: Technology at the limits of comprehension*. New York: Penguin.
- Aristoteles (2021). *Physik. Teilband 1: Bücher I bis IV*. Hrsg. und übers. von Gottfried Heinemann. Hamburg: Meiner. DOI: 10.28937/978-3-7873-3801-6.
- Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila & Francisco Herrera (2020). „Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI“. In: *Information Fusion*, Bd. 58, S. 82–115. DOI: 10.1016/j.inffus.2019.12.012.
- Avigad, Jeremy (2008). „Understanding Proofs“. In: *The Philosophy of Mathematical Practice*. Hrsg. von Paolo Mancosu. Oxford, New York: Oxford University Press, S. 317–353. DOI: 10.1093/acprof:oso/9780199296453.003.0013.
- Bach, Sebastian, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek (2015). „On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation“. In: *PLOS ONE*, Bd. 10, Nr. 7, Art. Nr. e0130140, S. 1–46. DOI: 10.1371/journal.pone.0130140.
- Bachmann, Marcus (2020). „The Epistemology of Understanding. A Contextualist Approach“. In: *Kriterion – Journal of Philosophy*, Bd. 34, Nr. 1, S. 75–98. DOI: 10.1515/krt-2020-340105.
- Baker, Gordon Park & Peter Michael Stephan Hacker, Hrsg. (2005). *Wittgenstein: Understanding and Meaning, Part I: Essays*. 2. Aufl. An Analytical Commentary on the *Philosophical Investigations* 1. Malden, Oxford, Carlton: Blackwell. DOI: 10.1002/9780470752807.
- Barman, Kristian Gonzalez, Sascha Caron, Tom Claassen & Henk de Regt (2024). „Towards a Benchmark for Scientific Understanding in Humans and Machines“. In: *Minds and Machines*, Bd. 34, Nr. 1, Art. Nr. 6, S. 1–16. DOI: 10.1007/s11023-024-09657-1.
- Bau, David, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou & Antonio Torralba (2020). „Understanding the Role of Individual Units in a Deep Neural Network“. In: *Proceedings of the National Academy of Sciences*, Bd. 117, Nr. 48, S. 30071–30078. DOI: 10.1073/pnas.1907375117.
- Baum, Kevin, Susanne Mantel, Eva Schmidt & Timo Speith (2022). „From Responsibility to Reason-Giving Explainable Artificial Intelligence“. In: *Philosophy & Technology*, Bd. 35, Nr. 1, Art. Nr. 12, S. 1–30. DOI: 10.1007/s13347-022-00510-w.
- Baumberger, Christoph (2011). „Types of Understanding: Their Nature and Their Relation to Knowledge“. In: *Conceptus*, Bd. 40, Nr. 98, S. 67–88. DOI: 10.1515/cpt-2014-0002.
- (2019). „Explicating Objectual Understanding: Taking Degrees Seriously“. In: *Journal for General Philosophy of Science*, Bd. 50, Nr. 3, S. 367–388. DOI: 10.1007/s10838-019-09474-6.
- Baumberger, Christoph, Claus Beisbart & Georg Brun (2017). „What Is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 1–34.

- Baumberger, Christoph & Georg Brun (2017). „Dimensions of Objectual Understanding“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 165–189.
- (2021). „Reflective Equilibrium and Understanding“. In: *Synthese*, Bd. 198, Nr. 8, S. 7923–7947. DOI: 10.1007/s11229-020-02556-9.
- BBC (1. Juli 2015). *Google apologises for Photos app’s racist blunder*. BBC News. URL: <https://www.bbc.com/news/technology-33347866> (besucht am 05. 08. 2026).
- Bechtel, William (2011). „Mechanism and Biological Explanation“. In: *Philosophy of Science*, Bd. 78, Nr. 4, S. 533–557. DOI: 10.1086/661513.
- Bechtel, William & Adele Abrahamsen (2005). „Explanation: A Mechanist Alternative“. In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, Bd. 36, Nr. 2, S. 421–441. DOI: 10.1016/j.shpsc.2005.03.010.
- Beierle, Christoph & Gabriele Kern-Isberner (2019). *Methoden wissenschaftlicher Systeme. Grundlagen, Algorithmen, Anwendungen*. 6. Aufl. Wiesbaden: Springer Vieweg. DOI: 10.1007/978-3-658-27084-1.
- Belkoniene, Miloud (2022). „The Rational Dimension of Understanding“. In: *Synthese*, Bd. 200, Nr. 5, Art. Nr. 349, S. 1–16. DOI: 10.1007/s11229-022-03839-z.
- (2023). *Rational Understanding: From Explanation to Knowledge*. Routledge Studies in Epistemology. New York, Oxon: Routledge. DOI: 10.4324/9781003324263.
- Belle, Vaishak & Ioannis Papantonis (2021). „Principles and Practice of Explainable Machine Learning“. In: *Frontiers in Big Data*, Bd. 4, Art. Nr. 688969, S. 1–25. DOI: 10.3389/fdata.2021.688969.
- Bengson, John (2018). „The Unity of Understanding“. In: *Making Sense of the World: New Essays on the Philosophy of Understanding*. Hrsg. von Stephen R. Grimm. Oxford, New York: Oxford University Press, S. 14–53. DOI: 10.1093/oso/9780190469863.003.0002.
- Berkeley, George (1949). *A Treatise Concerning the Principles of Human Knowledge*. Hrsg. von Arthur Aston Luce & Thomas Edmund Jessop. The Works of George Berkeley, Bishop of Cloyne 2. London: Thomas Nelson & Sons.
- Bien, Jacob & Robert Tibshirani (2011). „Prototype Selection for Interpretable Classification“. In: *The Annals of Applied Statistics*, Bd. 5, Nr. 4, S. 2403–2424. DOI: 10.1214/11-aos495.
- Bloom, Benjamin S., Max D. Engelhart, Edward J. Furst, Walker H. Hill & David R. Krathwohl (1956). *Taxonomy of Educational Objectives. The Classification of Educational Goals. Handbook 1: Cognitive Domain*. London: Longmans.
- Boden, Margaret A. (2014). „GOFAI“. In: *The Cambridge Handbook of Artificial Intelligence*. Hrsg. von Keith Frankish & William M. Ramsey. Cambridge: Cambridge University Press, S. 89–107. DOI: 10.1017/cbo9781139046855.007.
- Boge, Florian J. (2024). „Functional Concept Proxies and the Actually Smart Hans Problem: What’s Special about Deep Neural Networks in Science“. In: *Synthese*, Bd. 203, Nr. 1, Art. Nr. 16, S. 1–39. DOI: 10.1007/s11229-023-04440-8.

- Bokulich, Alisa (2016). „Fiction As a Vehicle for Truth: Moving beyond the Ontic Conception“. In: *The Monist*, Bd. 99, Nr. 3, S. 260–279. DOI: 10.1093/monist/onw004.
- BonJour, Laurence (1985). *The Structure of Empirical Knowledge*. Cambridge: Harvard University Press.
- Bordt, Sebastian, Michèle Finck, Eric Raidl & Ulrike von Luxburg (2022). „Post-Hoc Explanations Fail to Achieve Their Purpose in Adversarial Contexts“. In: *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (Seoul, 21.–24. Juni 2022). New York: Association for Computing Machinery (ACM), S. 891–905. DOI: 10.1145/3531146.3533153.
- Bourget, David (2017). „The Role of Consciousness in Grasping and Understanding“. In: *Philosophy and Phenomenological Research*, Bd. 95, Nr. 2, S. 285–318. DOI: 10.1111/phpr.12208.
- (2024). „A Phenomenal Theory of Grasping and Understanding“. In: *Understanding and Conscious Experience: Philosophical and Scientific Perspectives*. Hrsg. von Andrei Ionuț Mărașoiu & Mircea Dumitru. Routledge Studies in the Philosophy of Science. New York, Oxon: Routledge, S. 99–132. DOI: 10.4324/9781003437239-8.
- Boyd, Kenneth (2017). „Testifying Understanding“. In: *Episteme*, Bd. 14, Nr. 1, S. 103–127. DOI: 10.1017/epi.2015.53.
- Brogaard, Berit (2005). „I Know. Therefore, I Understand“. Manuskript.
- (2009). „What Mary Did Yesterday: Reflections on Knowledge-Wh“. In: *Philosophy and Phenomenological Research*, Bd. 78, Nr. 2, S. 439–467. DOI: 10.1111/j.1933-1592.2009.00250.x.
- Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi & Justin Gilmer (2017). „Adversarial Patch“. arXiv: 1712.09665.
- Brun, Georg (2016). „Explication As a Method of Conceptual Re-Engineering“. In: *Erkenntnis*, Bd. 81, Nr. 6, S. 1211–1241. DOI: 10.1007/s10670-015-9791-5.
- Buchanan, Bruce G., Randall Davis, Reid G. Smith & Edward A. Feigenbaum (2018). „Expert Systems: A Perspective from Computer Science“. In: *The Cambridge Handbook of Expertise and Expert Performance*. Hrsg. von K. Anders Ericsson, Robert R. Hoffman, Aaron Kozbelt & A. Mark Williams. 2. Aufl. Cambridge, New York, Melbourne, Neu Delhi, Singapur: Cambridge University Press, S. 84–104. DOI: 10.1017/9781316480748.007.
- Buckner, Cameron Joseph (2018). „Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks“. In: *Synthese*, Bd. 195, Nr. 12, S. 5339–5372. DOI: 10.1007/s11229-018-01949-1.
- (2019). „Deep Learning: A Philosophical Introduction“. In: *Philosophy Compass*, Bd. 14, Nr. 10, Art. Nr. e12625, S. 1–19. DOI: 10.1111/phc3.12625.
- Buijsman, Stefan (2026). „Machine Learning Models as Mathematics: Interpreting Explainable AI in Non-causal Terms“. In: *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Hrsg. von Juan Durán & Giorgia Pozzi. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science. Cham: Springer, S. 157–173. DOI: 10.1007/978-3-032-03083-2_8.

- Burkart, Nadia & Marco F. Huber (2021). „A Survey on the Explainability of Supervised Machine Learning“. In: *Journal of Artificial Intelligence Research*, Bd. 70, S. 245–317. DOI: 10.1613/jair.1.12228.
- Burrell, Jenna (2016). „How the Machine ‘thinks’: Understanding Opacity in Machine Learning Algorithms“. In: *Big Data & Society*, Bd. 3, Nr. 1, S. 1–12. DOI: 10.1177/2053951715622512.
- Callahan, Laura Frances (2018). „Moral Testimony: A Re-Conceived Understanding Explanation“. In: *The Philosophical Quarterly*, Bd. 68, Nr. 272, S. 437–459. DOI: 10.1093/pq/pqx057.
- Cammarata, Nick, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov & Ludwig Schubert (2020a). „Thread: Circuits“. In: *Distill*, Bd. 5, Nr. 3. DOI: 10.23915/distill.00024.
- Cammarata, Nick, Gabriel Goh, Shan Carter, Ludwig Schubert, Michael Petrov & Chris Olah (2020b). „Curve Detectors“. In: *Distill*, Bd. 5, Nr. 6. DOI: 10.23915/distill.00024.003.
- Cappelen, Herman (2018). *Fixing Language: An Essay on Conceptual Engineering*. Oxford, New York: Oxford University Press. DOI: 10.1093/oso/9780198814719.001.0001.
- Carnap, Rudolf (1971). *Logical Foundations of Probability*. 4. Aufl. London: Routledge & Kegan Paul.
- Carroll, Lewis (1894). *Sylvie and Bruno Concluded*. New York: Macmillan.
- Carter, J. Adam (2014). „Relativism, Knowledge and Understanding“. In: *Episteme*, Bd. 11, Nr. 1, S. 35–52. DOI: 10.1017/epi.2013.45.
- Carter, J. Adam & Emma C. Gordon (2011). „Norms of Assertion: The Quantity and Quality of Epistemic Support“. In: *Philosophia*, Bd. 39, Nr. 4, S. 615–635. DOI: 10.1007/s11406-011-9317-6.
- (2014). „Objectual Understanding and the Value Problem“. In: *American Philosophical Quarterly*, Bd. 51, Nr. 1, S. 1–13.
- (2016). „Objectual Understanding, Factivity and Belief“. In: *Epistemic Reasons, Norms and Goals*. Hrsg. von Martin Grajner & Pedro Schmechtig. Berlin, Boston: De Gruyter, S. 423–442. DOI: 10.1515/9783110496765-020.
- Chalmers, David J. (2020). „What Is Conceptual Engineering and What Should It Be?“ In: *Inquiry*, S. 1–18. DOI: 10.1080/0020174x.2020.1817141.
- Chisholm, Roderick Milton (2008). „The Myth of the Given“. In: *Epistemology: An Anthology*. Hrsg. von Ernest Sosa, Jaegwon Kim, Jeremy Fantl & Matthew McGrath. 2. Aufl. Malden, Oxford, Carlton: Blackwell, S. 80–93.
- Chollet, François (2021). *Deep Learning with Python*. 2. Aufl. Shelter Island: Manning.
- Clancey, William J. (1983). „The Epistemology of a Rule-Based Expert System—A Framework for Explanation“. In: *Artificial Intelligence*, Bd. 20, Nr. 3, S. 215–251. DOI: 10.1016/0004-3702(83)90008-5.
- Cohen, L. Jonathan (1995). *An Essay on Belief and Acceptance*. Oxford, New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780198236047.001.0001.
- Conee, Earl & Richard Feldman (2001). „Internalism Defended“. In: *Epistemology: Internalism and Externalism*. Hrsg. von Hilary Kornblith. Blackwell Readings in Philosophy. Malden, Oxford: Blackwell, S. 231–260.

- (2004). *Evidentialism*. Oxford, New York: Oxford University Press. DOI: 10.1093/0199253722.001.0001.
- Cooper, Neil (1994). „Understanding“. In: *Aristotelian Society Supplementary Volume*, Bd. 68, Nr. 1, S. 1–26. DOI: 10.1093/aristoteliansupp/68.1.1.
- (1995). „The Epistemology of Understanding“. In: *Inquiry*, Bd. 38, Nr. 3, S. 205–215. DOI: 10.1080/00201749508602386.
- Crane, Tim (2009). „Is Perception a Propositional Attitude?“. In: *The Philosophical Quarterly*, Bd. 59, Nr. 236, S. 452–469. DOI: 10.1111/j.1467-9213.2008.608.x.
- Craver, Carl F. (2014). „The Ontic Account of Scientific Explanation“. In: *Explanation in the Special Sciences: The Case of Biology and History*. Hrsg. von Marie I. Kaiser, Oliver R. Scholz, Daniel Plenge & Andreas Hüttemann. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 367. Dordrecht: Springer, S. 27–52. DOI: 10.1007/978-94-007-7563-3_2.
- Craver, Carl F. & Lindley Darden (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. Chicago, London: University of Chicago Press.
- Creel, Kathleen A. (2020). „Transparency in Complex Computational Systems“. In: *Philosophy of Science*, Bd. 87, Nr. 4, S. 568–589. DOI: 10.1086/709729.
- Crook, Barnaby, Maximilian Schlüter & Timo Speith (2023). „Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)“. In: *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops (REW)* (Hannover, 4.–8. Sep. 2023). Hrsg. von Kurt Schneider, Fabiano Dalpiaz & Jennifer Horkoff. Los Alamitos, Washington, Tokio: Institute of Electrical and Electronics Engineers (IEEE), S. 316–324. DOI: 10.1109/rew57809.2023.00060.
- Dam, Hoa Khanh, Truyen Tran & Aditya Ghose (2018). „Explainable Software Analytics“. In: *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)* (Gothenburg, 27. Mai–3. Juni 2018). New York: Association for Computing Machinery (ACM), S. 53–56. DOI: 10.1145/3183399.3183424.
- Dastin, Jeffrey (11. Okt. 2018). *Insight – Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. URL: <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/> (besucht am 05. 08. 2026).
- Davidsson, Paul, Franziska Klügl & Harko Verhagen (2017). „Simulation of Complex Systems“. In: *Springer Handbook of Model-Based Science*. Hrsg. von Lorenzo Magnani & Tommaso Bertolotti. Dordrecht, Heidelberg, London, New York: Springer, S. 783–797. DOI: 10.1007/978-3-319-30526-4_35.
- de Graaf, Maartje M. A. & Bertram F. Malle (2017). *How People Explain Action (and Autonomous Intelligent Systems Should Too)*. Techn. Ber. Artificial Intelligence for Human-Robot Interaction. Association for the Advancement of Artificial Intelligence (AAAI), FS-17-01, S. 19–26.

- de Regt, Henk W. (2009). „Understanding and Scientific Explanation“. In: *Scientific Understanding: Philosophical Perspectives*. Hrsg. von Henk W. de Regt, Sabina Leonelli & Kai Eigner. Pittsburgh: University of Pittsburgh Press, S. 21–42.
- (2015). „Scientific Understanding: Truth or Dare?“ In: *Synthese*, Bd. 192, Nr. 12, S. 3781–3797. DOI: 10.1007/s11229-014-0538-7.
- (2017). *Understanding Scientific Understanding*. Oxford, New York: Oxford University Press. DOI: 10.1093/oso/9780190652913.001.0001.
- de Regt, Henk W. & Dennis Dieks (2005). „A Contextual Approach to Scientific Understanding“. In: *Synthese*, Bd. 144, Nr. 1, S. 137–170. DOI: 10.1007/s11229-005-5000-4.
- de Regt, Henk W. & Victor Gijsbers (2017). „How False Theories Can Yield Genuine Understanding“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 50–75.
- Deck, Luca, Astrid Schomäcker, Timo Speith, Jakob Schöffner, Lena Kästner & Niklas Kühl (2024). „Mapping the Potential of Explainable AI for Fairness along the AI Lifecycle“. In: *Proceedings of the 3rd European Workshop on Algorithmic Fairness (EWAf)* (Mainz, 1.–3. Juli 2024). Hrsg. von Mattia Cerrato, Alesia Vallenias Coronel, Petra Ahrweiler, Michele Loi, Mykola Pechenizkiy & Aurelia Tamò-Larrieux. CEUR Workshop Proceedings 3908. Aachen.
- Delarivière, Sven & Bart Van Kerkhove (2021). „The Mark of Understanding: In Defense of an Ability Account“. In: *Axiomathes*, Bd. 31, Nr. 5, S. 619–648. DOI: 10.1007/s10516-020-09529-0.
- Dellsén, Finnur (2017). „Understanding without Justification or Belief“. In: *Ratio*, Bd. 30, Nr. 3, S. 239–254. DOI: 10.1111/rati.12134.
- (2018). „Deductive Cogency, Understanding, and Acceptance“. In: *Synthese*, Bd. 195, Nr. 7, S. 3121–3141. DOI: 10.1007/s11229-017-1365-4.
- (2020). „Beyond Explanation: Understanding As Dependency Modelling“. In: *The British Journal for the Philosophy of Science*, Bd. 71, Nr. 4, S. 1261–1286. DOI: 10.1093/bjps/axy058.
- Dieterich, William, Christina Mendoza & Tim Brennan (8. Juli 2016). *Compas Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Forschungsber. Northpointe Inc. Research Department.
- Duncan, Matt (2023). „Suppose We Know Things“. In: *Episteme*, Bd. 20, Nr. 2, S. 308–323. DOI: 10.1017/epi.2021.29.
- Durán, Juan Manuel & Nico Formanek (2018). „Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism“. In: *Minds and Machines*, Bd. 28, Nr. 4, S. 645–666. DOI: 10.1007/s11023-018-9481-6.
- Durán, Juan Manuel & Karin Rolanda Jongsma (2021). „Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI“. In: *Journal of Medical Ethics*, Bd. 47, Nr. 5, S. 329–335. DOI: 10.1136/medethics-2020-106820.
- Durán, Juan Manuel & Giorgia Pozzi (2025). „Trust and Trustworthiness in AI“. In: *Philosophy & Technology*, Bd. 38, Nr. 1, Art. Nr. 16, S. 1–31. DOI: 10.1007/s13347-025-00843-2.

- Dutilh Novaes, Catarina (2020). „Carnapian Explication and Ameliorative Analysis: A Systematic Comparison“. In: *Synthese*, Bd. 197, Nr. 3, S. 1011–1034. DOI: 10.1007/s11229-018-1732-9.
- Eaton, Marcia (1980). „Truth in Pictures“. In: *The Journal of Aesthetics and Art Criticism*, Bd. 39, Nr. 1, S. 15–26. DOI: 10.2307/429915.
- Egler, Miguel (2021). „Why Understanding-Why Is Contrastive“. In: *Synthese*, Bd. 199, Nr. 3–4, S. 6061–6083. DOI: 10.1007/s11229-021-03059-x.
- Elgin, Catherine Z. (2006). „From Knowledge to Understanding“. In: *Epistemology Futures*. Hrsg. von Stephen Hetherington. Oxford: Clarendon Press, S. 199–215.
- (2007). „Understanding and the Facts“. In: *Philosophical Studies*, Bd. 132, Nr. 1, S. 33–42. DOI: 10.1007/s11098-006-9054-z.
- (2014). „Non-Foundationalist Epistemology: Holism, Coherence, and Tenability“. In: *Contemporary Debates in Epistemology*. Hrsg. von Matthias Steup, John Turri & Ernest Sosa. 2. Aufl. Malden, Oxford, Chichester: Blackwell, S. 244–255.
- (2017). *True Enough*. Cambridge: MIT Press.
- (2021). „Replies“. In: *Synthese*, Bd. 199, Nr. 1-2, S. 1577–1597. DOI: 10.1007/s11229-020-02819-5.
- (2022). „Models As Felicitous Falsehoods“. In: *Principia: An International Journal of Epistemology*, Bd. 26, Nr. 1, S. 7–23. DOI: 10.5007/1808-1711.2022.e84576.
- Elsayed, Gamaleldin, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein (2019). „Adversarial Examples that Fool both Computer Vision and Time-Limited Humans“. In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)* (Montreal, 2.–8. Dez. 2018). Hrsg. von Samy Bengio, Hanna Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi & Roman Garnett. Bd. 1. Advances in Neural Information Processing Systems 31. Red Hook: Curran Associates, Inc., S. 3910–3920.
- Endsley, Mica R. (1995). „Toward a Theory of Situation Awareness in Dynamic Systems“. In: *Human Factors*, Bd. 37, Nr. 1, S. 32–64. DOI: 10.1518/001872095779049543.
- Erasmus, Adrian, Tyler D. P. Brunet & Eyal Fisher (2021). „What Is Interpretability?“ In: *Philosophy & Technology*, Bd. 34, S. 833–862. DOI: 10.1007/s13347-020-00435-2.
- Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun (2017). „Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks“. In: *Nature*, Bd. 542, Nr. 7639, S. 115–118. DOI: 10.1038/nature21056.
- Europäisches Parlament und Rat der Europäischen Union (27. Apr. 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)“. In: *Amtsblatt der Europäischen Union*. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj> (besucht am 28.08.2026).
- (13. Juni 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)“. In: *Amtsblatt der Europäischen Union*. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> (besucht am 28. 08. 2026).
- Evolving Artificial Intelligence Lab (30. Juni 2020). *Synthesizing: Code for paper “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”*. Lizenziert via MIT License. GitHub. URL: <https://github.com/Evolving-AI-Lab/synthesizing> (besucht am 05. 08. 2026).
- Facchini, Alessandro & Alberto Termine (2022). „Towards a Taxonomy for the Opacity of AI Systems“. In: *Proceedings of the 4th Conference on Philosophy and Theory of Artificial Intelligence (PT-AI)* (Gothenburg, 27.–28. Sep. 2021). Hrsg. von Vincent C. Müller. *Studies in Applied Philosophy, Epistemology and Rational Ethics* 63. Cham: Springer, S. 73–89. DOI: 10.1007/978-3-031-09153-7_7.
- Fantl, Jeremy & Matthew McGrath (2009). *Knowledge in an Uncertain World*. Oxford, New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780199550623.001.0001.
- Fara, Michael (2008). „Masked Abilities and Compatibilism“. In: *Mind*, Bd. 117, Nr. 468, S. 843–865. DOI: 10.1093/mind/fzn078.
- Farkas, Katalin (2019). „Objectual Knowledge“. In: *Acquaintance: New Essays*. Hrsg. von Jonathan Knowles & Thomas Raleigh. Oxford, New York: Oxford University Press, S. 260–276. DOI: 10.1093/oso/9780198803461.003.0013.
- Feldman, Richard (2003). *Epistemology*. Foundations of Philosophy. Upper Saddle River: Prentice Hall.
- Fleisher, Will (2022). „Understanding, Idealization, and Explainable AI“. In: *Episteme*, Bd. 19, Nr. 4, S. 534–560. DOI: 10.1017/epi.2022.39.
- Fong, Ruth C. & Andrea Vedaldi (2018). „Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks“. In: *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, 18.–22. Juni 2018). New York: Institute of Electrical and Electronics Engineers (IEEE), S. 8730–8738. DOI: 10.1109/cvpr.2018.00910.
- Frege, Gottlob (1892). „Über Sinn und Bedeutung“. In: *Zeitschrift für Philosophie und Philosophische Kritik*, Bd. 100, Nr. 1, S. 25–50.
- Freiesleben, Timo (2026). „Artificial Neural Nets and the Representation of Human Concepts“. In: *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Hrsg. von Juan Durán & Giorgia Pozzi. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 527. Cham: Springer, S. 273–294. DOI: 10.1007/978-3-032-03083-2_13.
- Freiesleben, Timo & Gunnar König (2023). „Dear XAI Community, We Need to Talk! Fundamental Misconceptions in Current XAI Research“. In: *Proceedings of the 1st World Conference on Explainable Artificial Intelligence (xAI)* (Lissabon, 26.–28. Juli 2023). Hrsg.

- von Luca Longo. Bd. 1. Communications in Computer and Information Science 1901. Cham: Springer, S. 48–65. DOI: 10.1007/978-3-031-44064-9_3.
- Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge & Felix A. Wichmann (2020). „Shortcut Learning in Deep Neural Networks“. In: *Nature Machine Intelligence*, Bd. 2, Nr. 11, S. 665–673. DOI: 10.1038/s42256-020-00257-z.
- Gerlings, Julie, Millie Søndergaard Jensen & Arisa Shollo (2022). „Explainable AI, but Explainable to Whom? An Exploratory Case Study of xAI in Healthcare“. In: *Handbook of Artificial Intelligence in Healthcare. Practicalities and Prospects*. Hrsg. von Chee-Peng Lim, Yen-Wei Chen, Ashlesha Vaidya, Charu Mahorkar & Lakhmi C. Jain. Bd. 2. Intelligent Systems Reference Library 212. Cham: Springer, S. 169–198. DOI: 10.1007/978-3-030-83620-7_7.
- Gettier, Edmund L. (1963). „Is Justified True Belief Knowledge?“ In: *Analysis*, Bd. 23, Nr. 6, S. 121–123. DOI: 10.1093/analys/23.6.121.
- Ghassemi, Marzyeh, Luke Oakden-Rayner & Andrew L. Beam (2021). „The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care“. In: *The Lancet Digital Health*, Bd. 3, Nr. 11, S. 745–750. DOI: 10.1016/s2589-7500(21)00208-9.
- Gichoya, Judy Wawira, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman & Haoran Zhang (2022). „AI Recognition of Patient Race in Medical Imaging: A Modelling Study“. In: *The Lancet Digital Health*, Bd. 4, Nr. 6, S. 406–414. DOI: 10.1016/s2589-7500(22)00063-2.
- Gijsbers, Victor (2013). „Understanding, Explanation, and Unification“. In: *Studies in History and Philosophy of Science Part A*, Bd. 44, Nr. 3, S. 516–522. DOI: 10.1016/j.shpsa.2012.12.003.
- Gildenblat, Jacob (10. Juli 2026). *PyTorch Grad-CAM: Advanced AI Explainability for computer vision. Support for CNNs, Vision Transformers, Classification, Object detection, Segmentation, Image similarity and more*. Lizenziert via MIT License. GitHub. URL: <https://github.com/jacobgil/pytorch-grad-cam> (besucht am 05. 08. 2026).
- Goldman, Alvin Ira (1979). „What Is Justified Belief?“ In: *Justification and Knowledge: New Studies in Epistemology*. Hrsg. von George S. Pappas. Philosophical Studies Series in Philosophy 17. Dordrecht: D. Reidel, S. 1–23.
- (2018). „Expertise“. In: *Topoi*, Bd. 37, Nr. 1, S. 3–10. DOI: 10.1007/s11245-016-9410-3.
- Gombrich, Ernst H. J. (1972). „The Visual Image“. In: *Scientific American*, Bd. 227, Nr. 3, S. 82–96. DOI: 10.1038/scientificamerican0972-82.
- Goodfellow, Ian, Yoshua Bengio & Aaron Courville (2016). *Deep Learning*. Cambridge, MA: MIT Press.
- Goodman, Bryce & Seth Flaxman (2017). „European Union Regulations on Algorithmic Decision-Making and a ‘Right to Explanation’“. In: *AI Magazine*, Bd. 38, Nr. 3, S. 50–57. DOI: 10.1609/aimag.v38i3.2741.

- Goodman, Nelson & Catherine Z. Elgin (1993). *Revisionen. Philosophie und andere Künste und Wissenschaften*. Übers. von Bernd Philippi. Suhrkamp-Taschenbuch Wissenschaft 1050. Frankfurt am Main: Suhrkamp.
- Gordon, Emma C. (2012). „Is There Propositional Understanding?“ In: *Logos & Episteme*, Bd. 3, Nr. 2, S. 181–192. doi: 10.5840/logos-episteme20123234.
- (2017). „Social Epistemology and the Acquisition of Understanding“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 293–317.
- (2021). „Elgin on Understanding: How Does It Involve Know-How, Endorsement and Factivity?“ In: *Synthese*, Bd. 198, Nr. 6, S. 4955–4972. doi: 10.1007/s11229-019-02381-9.
- (o. D.). „Understanding in Epistemology“. In: *Internet Encyclopedia of Philosophy*. Hrsg. von James Fieser & Bradley Dowden. URL: <https://iep.utm.edu/understa/>.
- Görz, Günther, Tanya Braun & Ute Schmid (2021). „Einleitung“. In: *Handbuch der Künstlichen Intelligenz*. Hrsg. von Günther Görz, Ute Schmid & Tanya Braun. 6. Aufl. Berlin, Boston: De Gruyter, S. 1–26.
- Greco, John (2014). „Episteme: Knowledge and Understanding“. In: *Virtues and Their Vices*. Hrsg. von Kevin Timpe & Craig A. Boyd. Oxford, New York: Oxford University Press, S. 284–301. doi: 10.1093/acprof:oso/9780199645541.003.0014.
- Greimann, Dirk (2007). „Regeln für das korrekte Explizieren von Begriffen“. In: *Zeitschrift für Philosophische Forschung*, Bd. 61, Nr. 3, S. 261–282. doi: 10.3196/004433007782418090.
- Grigorescu, Sorin, Bogdan Trasnea, Tiberiu Cocias & Gigel Macesanu (2020). „A Survey of Deep Learning Techniques for Autonomous Driving“. In: *Journal of Field Robotics*, Bd. 37, Nr. 3, S. 362–386. doi: 10.1002/rob.21918.
- Grimm, Stephen R. (2006). „Is Understanding a Species of Knowledge?“ In: *The British Journal for the Philosophy of Science*, Bd. 57, Nr. 3, S. 515–535. doi: 10.1093/bjps/axl015.
- (2010). „The Goal of Explanation“. In: *Studies in History and Philosophy of Science Part A*, Bd. 41, Nr. 4, S. 337–344. doi: 10.1016/j.shpsa.2010.10.006.
- (2011). „Understanding“. In: *Routledge Companion to Epistemology*. Hrsg. von Sven Bernecker & Duncan Pritchard. New York: Routledge, S. 84–94.
- (2014). „Understanding as Knowledge of Causes“. In: *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*. Hrsg. von Abrol Fairweather. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 366. Cham, Heidelberg, New York, Dordrecht, London: Springer, S. 329–345. doi: 10.1007/978-3-319-04672-3_19.
- (2017). „Understanding and Transparency“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 212–229.
- (2020). „Transmitting Understanding and Know-How“. In: *What the Ancients Offer to Contemporary Epistemology*. Hrsg. von Stephen Hetherington & Nicholas D. Smith. New York: Routledge, S. 124–139.

- Grote, Thomas, Konstantin Genin & Emily Sullivan (2024). „Reliability in Machine Learning“. In: *Philosophy Compass*, Bd. 19, Nr. 5, Art. Nr. e12974, S. 1–11. DOI: 10.1111/phc3.12974.
- Grundmann, Thomas (2017). *Analytische Einführung in die Erkenntnistheorie*. 2. Aufl. Berlin, Boston: De Gruyter. DOI: 10.1515/9783110211849.
- Grzankowski, Alex (2015). „Pictures Have Propositional Content“. In: *Review of Philosophy and Psychology*, Bd. 6, Nr. 1, S. 151–163. DOI: 10.1007/s13164-014-0217-0.
- Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti & Dino Pedreschi (2019). „A Survey of Methods for Explaining Black Box Models“. In: *ACM Computing Surveys*, Bd. 51, Nr. 5, Art. Nr. 93, S. 1–42. DOI: 10.1145/3236009.
- Gunning, David & David W. Aha (2019). „DARPA’s Explainable Artificial Intelligence Program“. In: *AI Magazine*, Bd. 40, Nr. 2, S. 44–58. DOI: 10.1609/aimag.v40i2.2850.
- Haack, Susan (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology*. Oxford: Blackwell.
- (2000). „A Foundherentist Theory of Empirical Justification“. In: *Epistemology: An Anthology*. Hrsg. von Ernest Sosa & Jaegwon Kim. Malden: Blackwell, S. 226–236.
- Hannon, Michael (2021). „Recent Work in the Epistemology of Understanding“. In: *American Philosophical Quarterly*, Bd. 58, Nr. 3, S. 269–290. DOI: 10.2307/48616060.
- Hardwig, John (1985). „Epistemic Dependence“. In: *The Journal of Philosophy*, Bd. 82, Nr. 7, S. 335–349. DOI: 10.2307/2026523.
- Harman, Gilbert (1986). *Change in View. Principles of Reasoning*. Cambridge: MIT Press.
- Hart, Herbert Lionel Adolphus (2008). „Postscript: Responsibility and Retribution“. In: *Punishment and Responsibility: Essays in the Philosophy of Law*. Mit einer Einl. von John Gardner. 2. Aufl. Oxford, New York: Oxford University Press, S. 210–237. DOI: 10.1093/acprof:oso/9780199534777.003.0009.
- Haslanger, Sally (2000). „Gender and Race: (What) Are They? (What) Do We Want Them to Be?“. In: *Noûs*, Bd. 34, Nr. 1, S. 31–55. DOI: 10.1111/0029-4624.00201.
- He, Kaiming (28. Okt. 2017). *Deep Residual Networks: Deep Residual Learning for Image Recognition*. Lizenziert via MIT License. GitHub. URL: <https://github.com/KaimingHe/deep-residual-networks> (besucht am 05. 08. 2026).
- Hempel, Carl Gustav (1965). „Aspects of Scientific Explanation“. In: *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: The Free Press, S. 331–496.
- Hempel, Carl Gustav & Paul Oppenheim (1948). „Studies in the Logic of Explanation“. In: *Philosophy of Science*, Bd. 15, Nr. 2, S. 135–175. DOI: 10.1086/286983.
- Hendrycks, Dan, Kevin Zhao, Steven Basart, Jacob Steinhardt & Dawn Song (2021). „Natural Adversarial Examples“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville/Online, 19.–25. Juni 2021). New York: Institute of Electrical and Electronics Engineers (IEEE), S. 15262–15271. DOI: 10.1109/CVPR46437.2021.01501.

- Herzog, Christian (2022). „On the Ethical and Epistemological Utility of Explicable AI in Medicine“. In: *Philosophy & Technology*, Bd. 35, Nr. 2, Art. Nr. 50, S. 1–31. DOI: 10.1007/s13347-022-00546-y.
- Hieronymi, Pamela (2011). „XIV-Reasons for Action“. In: *Proceedings of the Aristotelian Society*, Bd. 111. Part 3, S. 407–427. DOI: 10.1111/j.1467-9264.2011.00316.x.
- Hills, Alison (2016). „Understanding Why“. In: *Noûs*, Bd. 50, Nr. 4, S. 661–688. DOI: 10.1111/nous.12092.
- Hubert, Mario & Federica Isabella Malfatti (2023). „Towards Ideal Understanding“. In: *Ergo*, Bd. 10, Nr. Art. Nr. 21, S. 578–611. DOI: 10.3998/ergo.4651.
- Illari, Phyllis (2013). „Mechanistic Explanation: Integrating the Ontic and Epistemic“. In: *Erkenntnis*, Bd. 78, Nr. 2. Supplement, S. 237–255. DOI: 10.1007/s10670-013-9511-y.
- Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran & Aleksander Madry (2019). „Adversarial Examples Are Not Bugs, They Are Features“. In: *Proceedings of the 32nd Advances in Neural Information Processing Systems (NeurIPS)* (Vancouver, 8.–14. Dez. 2019). Hrsg. von Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox & Roman Garnett. Bd. 1. Advances in Neural Information Processing Systems 32. Red Hook: Curran Associates, Inc., S. 125–136.
- Isaac, Manuel Gustavo, Steffen Koch & Ryan Nefdt (2022). „Conceptual Engineering: A Road Map to Practice“. In: *Philosophy Compass*, Bd. 17, Nr. 10, Art. Nr. e12879, S. 1–15. DOI: 10.1111/phc3.12879.
- Jaster, Romy (2020a). *Agents’ Abilities*. Philosophical Analysis 83. Berlin, Boston: De Gruyter. DOI: 10.1515/9783110650464.
- (2020b). „Précis zu: *Agents’ Abilities*“. In: *Zeitschrift für Philosophische Forschung*, Bd. 74, Nr. 3, S. 443–447. DOI: 10.3196/004433020830038741.
- Jebeile, Julie (2018a). „Collaborative Practice, Epistemic Dependence and Opacity: The Case of Space Telescope Data Processing“. In: *Philosophia Scientiae*, Nr. 22-2, S. 59–78. DOI: 10.4000/philosophiascientiae.1483.
- (2018b). „Explaining with Simulations: Why Visual Representations Matter“. In: *Perspectives on Science*, Bd. 26, Nr. 2, S. 213–238. DOI: 10.1162/posc_a_00273.
- Karimi, Amir-Hossein, Bernhard Schölkopf & Isabel Valera (2021). „Algorithmic Recourse. From Counterfactual Explanations to Interventions“. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (Online/Kanada, 3.–10. März 2021). New York: Association for Computing Machinery (ACM), S. 353–362. DOI: 10.1145/3442188.3445899.
- Kästner, Lena & Lise Marie Andersen (2018). „Intervening into Mechanisms: Prospects and Challenges“. In: *Philosophy Compass*, Bd. 13, Nr. 11, Art. Nr. e12546, S. 1–15. DOI: 10.1111/phc3.12546.
- Kästner, Lena & Barnaby Crook (2024). „Explaining AI through Mechanistic Interpretability“. In: *European Journal for Philosophy of Science*, Bd. 14, Nr. 4, Art. Nr. 52, S. 1–25. DOI: 10.1007/s13194-024-00614-4.

- Kästner, Lena & Philipp Haueis (2021). „Discovering Patterns: On the Norms of Mechanistic Inquiry“. In: *Erkenntnis*, Bd. 86, Nr. 6, S. 1635–1660. DOI: 10.1007/s10670-019-00174-7.
- Kauppinen, Antti (2024). „How Emotions Grasp Value“. In: *Philosophical Issues. A Supplement to Noûs*, Bd. 34, Nr. 1. Hrsg. von Ernest Sosa, Jaakko Hirvelä, Antti Kauppinen & Maria Lasonen-Aarnio, S. 213–233. DOI: 10.1111/phs.12272.
- Kelp, Christoph (2015). „Understanding Phenomena“. In: *Synthese*, Bd. 192, Nr. 12, S. 3799–3816. DOI: 10.1007/s11229-014-0616-x.
- (2017). „Towards a Knowledge-Based Account of Understanding“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 251–271.
- (2021). *Inquiry, Knowledge, and Understanding*. Oxford, New York: Oxford University Press. DOI: 10.1093/oso/9780192896094.001.0001.
- Kenny, Anthony (2010). „Concepts, Brains, and Behaviour“. In: *Grazer Philosophische Studien*, Bd. 81, Nr. 1, S. 105–113. DOI: 10.1163/9789042030190_007.
- Khalifa, Kareem (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press. DOI: 10.1017/9781108164276.
- Kim, Been, Rajiv Khanna & Oluwasanmi Koyejo (2016). „Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability“. In: *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)* (Barcelona, 5.–10. Dez. 2016). Hrsg. von Daniel D. Lee, Ulrike von Luxburg, Roman Garnett, Masashi Sugiyama & Isabelle Guyon. Bd. 1. Advances in Neural Information Processing Systems 29. Red Hook: Curran Associates Inc., S. 2288–2296.
- Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas & Rory Sayres (2018). „Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)“. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)* (Stockholmsmässan, 10.–15. Juli 2018). Hrsg. von Jennifer Dy & Andreas Krause. Bd. 80. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research (PMLR), S. 2668–2677.
- Kim, Jaegwon (1994). „Explanatory Knowledge and Metaphysical Dependence“. In: *Philosophical Issues*, Bd. 5, S. 51–69. DOI: 10.2307/1522873.
- Kjeldsen, Jens E. (2015). „The Study of Visual and Multimodal Argumentation“. In: *Argumentation*, Bd. 29, Nr. 2, S. 115–132. DOI: 10.1007/s10503-015-9348-4.
- Kleinberg, Jon, Sendhil Mullainathan & Manish Raghavan (2017). „Inherent Trade-Offs in the Fair Determination of Risk Scores“. In: *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS)* (Berkeley). Hrsg. von Christos H. Papadimitriou. Leibniz International Proceedings in Informatics (LIPIcs) 67. Saarbrücken/Wadern: Leibniz-Zentrum für Informatik, Art. Nr. 43, S. 1–23. DOI: 10.4230/LIPICS.ITCS.2017.43.
- Knüsel, Benedikt & Christoph Baumberger (2020). „Understanding Climate Phenomena with Data-Driven Models“. In: *Studies in History and Philosophy of Science Part A*, Bd. 84, S. 46–56. DOI: 10.1016/j.shpsa.2020.08.003.

- Koch, Steffen (2022). „Conceptual Engineering: Begriffe auf dem Prüfstand“. In: *Anspruch und Methode der Philosophie. Stimmen aus der Gegenwart*. Hrsg. von Niklas Grouls & Laura Martena. Schriften zur Anthropologie und Phänomenologie 2. Darmstadt: wbg, S. 101–129.
- Koh, Pang Wei & Percy Liang (2018). „Understanding Black-Box Predictions via Influence Functions“. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)* (Sydney, 6.–11. Aug. 2017). Hrsg. von Doina Precup & Yee Whye Teh. Bd. 4. Proceedings of Machine Learning Research 70. Red Hook: Curran Associates, Inc., S. 2976–2987.
- Koppelberg, Dirk (2019). „Tugendbasierte Theorien des Wissens und Tugenderkenntnistheorie“. In: *Handbuch Erkenntnistheorie*. Hrsg. von Martin Grajner & Guido Melchior. Stuttgart: J.B. Metzler, S. 79–86.
- Korsmeyer, Carolyn (1985). „Pictorial Assertion“. In: *The Journal of Aesthetics and Art Criticism*, Bd. 43, Nr. 3, S. 257–265. DOI: 10.2307/430639.
- Kosso, Peter (2002). „The Omniscient: Beauty and Scientific Understanding“. In: *International Studies in the Philosophy of Science*, Bd. 16, Nr. 1, S. 39–48. DOI: 10.1080/02698590120118819.
- (2007). „Scientific Understanding“. In: *Foundations of Science*, Bd. 12, Nr. 2, S. 173–188. DOI: 10.1007/s10699-006-0002-3.
- Kozak, Piotr (2023). *Thinking in Images: Imagistic Cognition and Non-Propositional Content*. London: Bloomsbury. DOI: 10.5040/9781350267497.
- Krathwohl, David R. (2002). „A Revision of Bloom’s Taxonomy: An Overview“. In: *Theory Into Practice*, Bd. 41, Nr. 4, S. 212–218. DOI: 10.1207/s15430421tip4104_2.
- Krishnan, Maya (2020). „Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning“. In: *Philosophy & Technology*, Bd. 33, Nr. 3, S. 487–502. DOI: 10.1007/s13347-019-00372-9.
- Kuhn, Thomas S. (1977). „Objectivity, Value Judgment, and Theory Choice. Selected Studies in Scientific Tradition and Change“. In: *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago, London: University of Chicago Press, S. 320–339.
- Kulvicki, John (2020). *Modeling the Meanings of Pictures: Depiction and the Philosophy of Language*. Oxford, New York: Oxford University Press. DOI: 10.1093/oso/9780198847472.001.0001.
- Kuorikoski, Jaakko (2012). „Simulation and the Sense of Understanding“. In: *Models, Simulations, and Representations*. Hrsg. von Paul Humphreys & Cyrille Imbert. New York: Routledge, S. 186–205.
- Kurz, Constanze & Stefan Ullrich (7. Apr. 2025). *Automatisierte Rasterfahndung: Tür zu für Palantir und Co*. Lizenziert via CC BY-NC-SA 4.0 Namensnennung – Nicht kommerziell – Share Alike 4.0 International, <https://creativecommons.org/licenses/by-nc-sa/4.0/>. netzpolitik.org e. V. URL: <https://netzpolitik.org/2025/automatisierte-rasterfahndung-tuer-zu-fuer-palantir-und-co/> (besucht am 05. 08. 2026).

- Kvanvig, Jonathan L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge Studies in Philosophy. Cambridge, New York: Cambridge University Press. DOI: 10.1017/cbo9780511498909.
- (2009a). „Responses to Critics“. In: *Epistemic Value*. Hrsg. von Adrian Haddock, Alan Millar & Duncan Pritchard. Oxford, New York: Oxford University Press, S. 339–351. DOI: 10.1093/acprof:oso/9780199231188.005.0007.
 - (2009b). „The Value of Understanding“. In: *Epistemic Value*. Hrsg. von Adrian Haddock, Alan Millar & Duncan Pritchard. Oxford, New York: Oxford University Press, S. 95–111. DOI: 10.1093/acprof:oso/9780199231188.003.0005.
- Lackey, Jennifer (2008). *Learning from Words: Testimony as a Source of Knowledge*. Oxford, New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780199219162.001.0001.
- Ladyman, James, James Lambert & Karoline Wiesner (2013). „What Is a Complex System?“ In: *European Journal for Philosophy of Science*, Bd. 3, Nr. 1, S. 33–67. DOI: 10.1007/s13194-012-0056-8.
- Langer, Markus, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesting & Kevin Baum (2021). „What Do We Want from Explainable Artificial Intelligence (XAI)? A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research“. In: *Artificial Intelligence*, Bd. 296, Art. Nr. 103473, S. 1–24. DOI: 10.1016/j.artint.2021.103473.
- Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek & Klaus-Robert Müller (2019). „Unmasking Clever Hans Predictors and Assessing What Machines Really Learn“. In: *Nature Communications*, Bd. 10, Nr. 1, S. 1–8. DOI: 10.1038/s41467-019-08987-4.
- Law, John & Michael Lynch (1988). „Lists, Field Guides, and the Descriptive Organization of Seeing: Birdwatching as an Exemplary Observational Activity“. In: *Human Studies*, Bd. 11, Nr. 2-3, S. 271–303. DOI: 10.1007/bf00177306.
- Lawler, Insa (2021). „Scientific Understanding and Felicitous Legitimate Falsehoods“. In: *Synthese*, Bd. 198, Nr. 7, S. 6859–6887. DOI: 10.1007/s11229-019-02495-0.
- LeCun, Yann & Ishan Misra (4. März 2021). *Self-Supervised Learning: The Dark Matter of Intelligence*. Meta AI. URL: <https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/> (besucht am 10. 10. 2025).
- Leonardi, Paul M. (2012). „Materiality, Sociomateriality, and Socio-Technical Systems: What Do These Terms Mean? How Are They Different? Do We Need Them?“ In: *Materiality and Organizing: Social Interaction in a Technological World*. Hrsg. von Paul M. Leonardi, Bonnie A. Nardi & Jannis Kallinikos. Oxford, New York: Oxford University Press, S. 24–48. DOI: 10.1093/acprof:oso/9780199664054.003.0002.
- Lewis, David K. (1973). „Causation“. In: *The Journal of Philosophy*, Bd. 70, Nr. 17, S. 556–567. DOI: 10.2307/2025310.
- (1983). „Mad Pain and Martian Pain“. In: *Philosophical Papers*. Bd. 1. New York: Oxford University Press, S. 122–132. DOI: 10.1093/0195032047.003.0009.

- Lipton, Peter (1990). „Contrastive Explanation“. In: *Royal Institute of Philosophy Supplement*, Bd. 27, S. 247–266. DOI: 10.1017/s1358246100005130.
- (2009). „Understanding without Explanation“. In: *Scientific Understanding: Philosophical Perspectives*. Hrsg. von Henk W. de Regt, Sabina Leonelli & Kai Eigner. Pittsburgh: University of Pittsburgh Press, S. 43–63.
- Lipton, Zachary C. (2018). „The Mythos of Model Interpretability“. In: *Queue*, Bd. 16, Nr. 3, S. 31–57. DOI: 10.1145/3236386.3241340.
- Locke, John (1979). *An Essay Concerning Human Understanding*. Hrsg. von Peter H. Nidditch. Oxford: Clarendon Press.
- Longo, Luca, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith & Simone Stumpf (2024). „Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions“. In: *Information Fusion*, Bd. 106, Art. Nr. 102301, S. 1–22. DOI: 10.1016/j.inffus.2024.102301.
- Lundberg, Scott M. & Su-In Lee (2017). „A Unified Approach to Interpreting Model Predictions“. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)* (Long Beach, 4.–9. Dez. 2017). Hrsg. von Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach & Rob Fergus. Red Hook: Curran Associates, Inc., S. 4768–4777.
- Lynch, Michael Patrick (2018). „Understanding and Coming to Understand“. In: *Making Sense of the World: New Essays on the Philosophy of Understanding*. Hrsg. von Stephen R. Grimm. Oxford, New York: Oxford University Press, S. 194–208. DOI: 10.1093/oso/9780190469863.003.0010.
- Machamer, Peter, Lindley Darden & Carl F. Craver (2000). „Thinking about Mechanisms“. In: *Philosophy of Science*, Bd. 67, Nr. 1, S. 1–25. DOI: 10.1086/392759.
- Maier, Andreas, Christopher Syben, Tobias Lasser & Christian Riess (2019). „A Gentle Introduction to Deep Learning in Medical Image Processing“. In: *Zeitschrift für Medizinische Physik*, Bd. 29, Nr. 2: *Special Issue: Deep Learning in Medical Physics*, S. 86–101. DOI: 10.1016/j.zemedi.2018.12.003.
- Malfatti, Federica Isabella (2023). *Verstehen verstehen. Eine erkenntnistheoretische Untersuchung*. Hrsg. von Katja Crone, Johannes Haag & David Löwenstein. Theoria 5. Basel, Berlin: Schwabe.
- Mann, Sara (2025). „Understanding via Exemplification in XAI: How Explaining Image Classification Benefits from EXEMPLARS“. In: *AI & Society*, Bd. 40, Nr. 1. Lizenziert via CC BY 4.0 Namensnennung 4.0 International, <https://creativecommons.org/licenses/by/4.0/>, S. 37–52. DOI: 10.1007/s00146-023-01837-4.
- Mann, Sara, Barnaby Crook, Lena Kästner, Astrid Schomäcker & Timo Speith (2023). „Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy“. In: *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*

- (REW) (Hannover, 4.–8. Sep. 2023). Hrsg. von Kurt Schneider, Fabiano Dalpiaz & Jennifer Horkoff. Copyright ©2023, IEEE. Los Alamitos, Washington, Tokio: Institute of Electrical and Electronics Engineers (IEEE), S. 337–342. DOI: 10.1109/REW57809.2023.00063.
- Markus, Aniek F., Jan A. Kors & Peter R. Rijnbeek (2021). „The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies“. In: *Journal of Biomedical Informatics*, Bd. 113, Art. Nr. 103655, S. 1–11. DOI: 10.1016/j.jbi.2020.103655.
- McCain, Kevin (2016). *The Nature of Scientific Knowledge: An Explanatory Approach*. Cham: Springer. DOI: 10.1007/978-3-319-33405-9.
- McCrudden, Matthew T. & David N. Rapp (2017). „How Visual Displays Affect Cognitive Processing“. In: *Educational Psychology Review*, Bd. 29, Nr. 3, S. 623–639. DOI: 10.1007/s10648-015-9342-2.
- McSweeney, Michaela M. (2024). „Why Mary Left Her Room“. In: *Philosophy and Phenomenological Research*, Bd. 109, Nr. 1, S. 261–287. DOI: 10.1111/phpr.13034.
- Meyer, Bertrand (28. Okt. 2011). *John McCarthy*. Communications of the Association for Computing Machinery (ACM). URL: <https://perma.cc/49S8-3GM6> (besucht am 10. 10. 2025).
- Miller, Tim (2019). „Explanation in Artificial Intelligence: Insights from the Social Sciences“. In: *Artificial Intelligence*, Bd. 267, S. 1–38. DOI: 10.1016/j.artint.2018.07.007.
- Miller, Tim, Piers Howe & Liz Sonenberg (2017). „Explainable AI: Beware of Inmates Running the Asylum. Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences“. In: *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)* (Melbourne, 20. Aug. 2017). Hrsg. von David W. Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut & Peter Stone. International Joint Conferences on Artificial Intelligence (IJCAI), S. 36–42.
- Mitchell, Melanie & David C. Krakauer (2023). „The Debate Over Understanding in AI’s Large Language Models“. In: *Proceedings of the National Academy of Sciences*, Bd. 120, Nr. 13, Art. Nr. e2215907120, S. 1–5. DOI: 10.1073/pnas.2215907120.
- Mitchell, Tom (1997). *Machine Learning*. New York: McGraw-Hill.
- Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter & Luciano Floridi (2016). „The Ethics of Algorithms: Mapping the Debate“. In: *Big Data & Society*, Bd. 3, Nr. 2, S. 1–21. DOI: 10.1177/2053951716679679.
- Mittelstadt, Brent Daniel, Chris Russell & Sandra Wachter (2019). „Explaining Explanations in AI“. In: *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAccT)* (Atlanta, 29.–31. Jan. 2019). New York: Association for Computing Machinery (ACM), S. 279–288. DOI: 10.1145/3287560.3287574.
- Mizrahi, Moti (2012). „Idealizations and Scientific Understanding“. In: *Philosophical Studies*, Bd. 160, Nr. 2, S. 237–252. DOI: 10.1007/s11098-011-9716-3.
- Mohseni, Sina, Niloofar Zarei & Eric D. Ragan (2020). „A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems“. In: *ACM Transactions on Interactive Intelligent Systems*, Bd. 11, Nr. 3-4, Art. Nr. 24, S. 1–45. DOI: 10.1145/3387166.

- Molnar, Christoph (2022). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2. Aufl. URL: <https://christophm.github.io/interpretable-ml-book/>.
- Moravcsik, Julius M. E. (1979). „Understanding“. In: *dialectica*, Bd. 33, Nr. 3–4, S. 201–216. DOI: 10.1111/j.1746-8361.1979.tb00752.x.
- Mueller, Shane T. (2020). „Cognitive Anthropomorphism of AI: How Humans and Computers Classify Images“. In: *Ergonomics in Design: The Quarterly of Human Factors Applications*, Bd. 28, Nr. 3, S. 12–19. DOI: 10.1177/1064804620920870.
- Nado, Jennifer (2021). „Conceptual Engineering, Truth, and Efficacy“. In: *Synthese*, Bd. 198, Nr. 7. Supplement, S. 1507–1527. DOI: 10.1007/s11229-019-02096-x.
- National Center for Biotechnology Information (NCBI), Hrsg. (2025). *The Discovery of the Double Helix, 1951–1953*. The Francis Crick Papers. Bethesda: National Library of Medicine (US), National Center for Biotechnology Information. URL: <https://profiles.nlm.nih.gov/spotlight/sc/feature/doublehelix> (besucht am 26. 11. 2025).
- Nauta, Meike, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen & Christin Seifert (2023). „From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI“. In: *ACM Computing Surveys*, Bd. 55, Nr. 13s, Art. Nr. 295, S. 1–42. DOI: 10.1145/3583558.
- Newman, Mark P. (2014). „EMU and Inference: What the Explanatory Model of Scientific Understanding Ignores“. In: *European Journal for Philosophy of Science*, Bd. 4, Nr. 1, S. 55–74. DOI: 10.1007/s13194-013-0075-0.
- (2017a). „An Evidentialist Account of Explanatory Understanding“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 190–211.
- (2017b). „Theoretical Understanding in Science“. In: *The British Journal for the Philosophy of Science*, Bd. 68, Nr. 2, S. 571–595. DOI: 10.1093/bjps/axv041.
- Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox & Jeff Clune (2016). „Synthesizing the preferred inputs for neurons in neural networks via deep generator networks“. In: arXiv: 1605.09304.
- Obermeyer, Ziad, Brian Powers, Christine Vogeli & Sendhil Mullainathan (2019). „Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations“. In: *Science*, Bd. 366, Nr. 6464, S. 447–453. DOI: 10.1126/science.aax2342.
- Oddie, Graham & Gustavo Cevolani (2022). „Truthlikeness“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta & Uri Nodelman. Winter-Edition. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/win2022/entries/truthlikeness/>.
- Olah, Christopher, Alexander Mordvintsev & Ludwig Schubert (2017). „Feature Visualization“. In: *Distill*, Bd. 2, Nr. 11. DOI: 10.23915/distill.00007.
- Olah, Christopher, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye & Alexander Mordvintsev (2018). „The Building Blocks of Interpretability“. In: *Distill*, Bd. 3, Nr. 3. DOI: 10.23915/distill.00010.

- Páez, Andrés (2019). „The Pragmatic Turn in Explainable Artificial Intelligence (XAI)“. In: *Minds and Machines*, Bd. 29, Nr. 3, S. 441–459. DOI: 10.1007/s11023-019-09502-w.
- (2026). „Axe the X in XAI: A Plea for Understandable AI“. In: *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Hrsg. von Juan Durán & Giorgia Pozzi. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 527. Cham: Springer, S. 133–156. DOI: 10.1007/978-3-032-03083-2_7.
- Pappas, George S. (2023). „Internalist vs. Externalist Conceptions of Epistemic Justification“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta & Uri Nodelman. Frühlings-Edition. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/spr2023/entries/justep-intext/>.
- Park, Seungbae (2017). „Understanding without Justification and Belief?“. In: *Principia: An International Journal of Epistemology*, Bd. 21, Nr. 3, S. 379–389. DOI: 10.5007/1808-1711.2017v21n3p379.
- Pearl, Judea & Dana Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.
- Pelland, Jean Charles, Jasmin Trächtler & Helene Love (2024). „Practice Makes Human: Why We Can’t Understand Black-Box Artificial Intelligence“. In: *Wittgenstein and Artificial Intelligence, Volume II: Values and Governance*. Hrsg. von Brian Ball, Alice C. Helliwell & Alessandro Rossi. Anthem Studies in Wittgenstein. London: Anthem Press, S. 183–202.
- Perini, Laura T. (2005a). „Explanation in Two Dimensions: Diagrams and Biological Explanation“. In: *Biology & Philosophy*, Bd. 20, Nr. 2–3, S. 257–269. DOI: 10.1007/s10539-005-2562-y.
- (2005b). „The Truth in Pictures“. In: *Philosophy of Science*, Bd. 72, Nr. 1, S. 262–285. DOI: 10.1086/426852.
- Peterson, Roger Tory (1980). *A Field Guide to the Birds: Eastern and Central North America*. Boston: Houghton Mifflin Harcourt.
- Potochnik, Angela (2015). „The Diverse Aims of Science“. In: *Studies in History and Philosophy of Science Part A*, Bd. 53, S. 71–80. DOI: 10.1016/j.shpsa.2015.05.008.
- (2016). „Scientific Explanation: Putting Communication First“. In: *Philosophy of Science*, Bd. 83, Nr. 5, S. 721–732. DOI: 10.1086/687858.
- Prasetya, Yunus (2022). „ANNs and Unifying Explanations: Reply to Erasmus, Brunet, and Fisher“. In: *Philosophy & Technology*, Bd. 35, Nr. 2, Art. Nr. 43. DOI: 10.1007/s13347-022-00540-4.
- Preece, Alun, Dan Harborne, Dave Braines, Richard Tomsett & Supriyo Chakraborty (2018). „Stakeholders in Explainable AI“. arXiv: 1810.00184.
- Prinzing, Michael (2018). „The Revisionist’s Rubric: Conceptual Engineering and the Discontinuity Objection“. In: *Inquiry*, Bd. 61, Nr. 8, S. 854–880. DOI: 10.1080/0020174x.2017.1385522.
- Pritchard, Duncan (2008). „Knowing the Answer, Understanding and Epistemic Value“. In: *Grazer Philosophische Studien*, Bd. 77, Nr. 1, S. 325–339. DOI: 10.1163/18756735-90000852.

- Pritchard, Duncan (2009). „Knowledge, Understanding and Epistemic Value“. In: *Royal Institute of Philosophy Supplement*, Bd. 64, S. 19–43. doi: 10.1017/s1358246109000046.
- (2010). „Understanding“. In: *The Nature and Value of Knowledge: Three Investigations*. Hrsg. von Duncan Pritchard, Alan Millar & Adrian Haddock. Oxford, New York: Oxford University Press, S. 66–88. doi: 10.1093/acprof:oso/9780199586264.003.0004.
 - (2014). „Knowledge and Understanding“. In: *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*. Hrsg. von Abrol Fairweather. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 366. Cham, Heidelberg, New York, Dordrecht, London: Springer, S. 315–327. doi: 10.1007/978-3-319-04672-3_18.
- Rabold, Johannes, Hannah Deininger, Michael Siebers & Ute Schmid (2020a). „Enriching Visual with Verbal Explanations for Relational Concepts — Combining LIME with Aleph“. In: *Proceedings of the International Workshops of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD): Machine Learning and Knowledge Discovery in Databases* (Würzburg, 16.–20. Sep. 2019). Hrsg. von Peggy Cellier & Kurt Driessens. Bd. 1. Communications in Computer and Information Science 1167. Cham: Springer, S. 180–192. doi: 10.1007/978-3-030-43823-4_16.
- Rabold, Johannes, Gesina Schwalbe & Ute Schmid (2020b). „Expressive Explanations of DNNs by Combining Concept Analysis with ILP“. In: *Proceedings of the 43rd German Conference on AI: Advances in Artificial Intelligence (KI)* (Bamberg, 21.–25. Sep. 2020). Hrsg. von Ute Schmid, Franziska Klügl & Diedrich Wolter. Lecture Notes in Computer Science 12325. Cham: Springer, S. 148–162. doi: 10.1007/978-3-030-58285-2_11.
- Ratti, Emanuele & Mark Graves (2022). „Explainable Machine Learning Practices: Opening Another Black Box for Reliable Medical AI“. In: *AI and Ethics*, Bd. 2, Nr. 4, S. 801–814. doi: 10.1007/s43681-022-00141-z.
- Riaz, Amber (2015). „Moral Understanding and Knowledge“. In: *Philosophical Studies*, Bd. 172, Nr. 1, S. 113–128. doi: 10.1007/s11098-014-0328-6.
- Ribeiro, Marco Túlio (25. Juli 2024). *Lime: Explaining the predictions of any machine learning classifier*. Lizenziert via BSD 2-Clause “Simplified” License. GitHub. URL: <https://github.com/marcotcr/lime> (besucht am 05. 08. 2026).
- Ribeiro, Marco Túlio, Sameer Singh & Carlos Guestrin (2016). „“Why Should I Trust You?” Explaining the Predictions of Any Classifier“. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (San Francisco, 13.–17. Aug. 2016). New York: Association for Computing Machinery (ACM), S. 1135–1144. doi: 10.1145/2939672.2939778.
- (2018). „Anchors: High-Precision Model-Agnostic Explanations“. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (New Orleans, 2.–7. Feb. 2018). Bd. 32. Nr. 1. Palo Alto: AAAI Press, S. 1527–1535. doi: 10.1609/aaai.v32i1.11491.
- Ribera, Mireia & Agata Lapedriza (2019). „Can We Do Better Explanations? A Proposal of User-Centered Explainable AI“. In: *Joint Proceedings of the ACM Conference on Intelligent*

- User Interfaces Workshops (ACM IUI WS)* (Los Angeles, 20. März 2019). Hrsg. von Christoph Trattner, Denis Parra & Nathalie Riche. New York: Association for Computing Machinery (ACM).
- Riggs, Wayne D. (2003). „Understanding ‘Virtue’ and the Virtue of Understanding“. In: *Intellectual Virtue: Perspectives from Ethics and Epistemology*. Hrsg. von Michael DePaul & Linda Zagzebski. Oxford, New York: Oxford University Press, S. 203–226. DOI: 10.1093/acprof:oso/9780199252732.003.0010.
- (2009). „Understanding, Knowledge, and the Meno Requirement“. In: *Epistemic Value*. Hrsg. von Adrian Haddock, Alan Millar & Duncan Pritchard. Oxford, New York: Oxford University Press, S. 331–338.
- Rohwer, Yasha & Collin C. Rice (2013). „Hypothetical Pattern Idealization and Explanatory Models“. In: *Philosophy of Science*, Bd. 80, Nr. 3, S. 334–355. DOI: 10.1086/671399.
- Rozenblit, Leonid & Frank C. Keil (2002). „The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth“. In: *Cognitive Science*, Bd. 26, Nr. 5, S. 521–562. DOI: 10.1207/s15516709cog2605_1.
- Rudin, Cynthia (2019). „Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead“. In: *Nature Machine Intelligence*, Bd. 1, Nr. 5, S. 206–215. DOI: 10.1038/s42256-019-0048-x.
- Russell, Bertrand (1910). „Knowledge by Acquaintance and Knowledge by Description“. In: *Proceedings of the Aristotelian Society*, Bd. 11, S. 108–128.
- Russell, Stuart & Peter Norvig (2021). *Artificial Intelligence. A Modern Approach*. Hoboken: Pearson.
- Ryle, Gilbert (1949). *The Concept of Mind*. London: Hutchinson & Co.
- Saibene, Aurora, Michela Assale & Marta Giltri (2021). „Expert Systems: Definitions, Advantages and Issues in Medical Field Applications“. In: *Expert Systems with Applications*, Bd. 177, Art. Nr. 114900, S. 1–15. DOI: 10.1016/j.eswa.2021.114900.
- Saint-Exupéry, Antoine (2023). *Der kleine Prinz*. Übers. von Ulrich Bossier. Reclams Universal-Bibliothek 14389. Ditzingen: Reclam.
- Salmon, Wesley C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press. DOI: 10.1515/9780691221489.
- Samek, Wojciech, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders & Klaus-Robert Müller (2021). „Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications“. In: *Proceedings of the IEEE*, Bd. 109, Nr. 3, S. 247–278. DOI: 10.1109/JPROC.2021.3060483.
- Sand, George (1884). „DCXXXVII: À M. Armand Barbès, à la Haye. Nohant, 12 mai 1867“. In: *Correspondance 1812–1876*. Bd. 5. Paris: Calmann Lévy, S. 195–197.
- Schlicker, Nadine, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C. Hirsch & Markus Langer (2025). „How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM)“. In: *Computers in Human Behavior*, Bd. 170, Art. Nr. 108671, S. 1–19. DOI: 10.1016/j.chb.2025.108671.

- Schmidt, Eva (2022). „Wie können wir autonomen KI-Systemen vertrauen?“ In: *Gratwanderung Künstliche Intelligenz*. Hrsg. von Britta Konz, Karl-Heinrich Ostmeyer & Marcel Scholz. Stuttgart: Kohlhammer, S. 11–29.
- (2025). „Reasons, Attenuators, and Virtue: A Novel Account of Pragmatic Encroachment“. In: *Analytic Philosophy*, Bd. 66, Nr. 2, S. 159–180. doi: 10.1111/phib.12314.
 - (2026). „Stakes and Understanding the Decisions of Artificial Intelligent Systems“. In: *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Hrsg. von Juan Durán & Giorgia Pozzi. Synthese Library: Studies in Epistemology, Logic, Methodology, and Philosophy of Science 527. Cham: Springer, S. 221–242. doi: 10.1007/978-3-032-03083-2_11.
- Schmidt, Eva & Sara Mann (o. D.). „AI Trustworthiness as Competence and Character“. In: *Philosophical Issues. A Supplement to Nous*, Bd. 36, Nr. 1. Hrsg. von Ernest Sosa & Will Fleisher. Akzeptiert, in Vorbereitung.
- Schmidt, Eva, Paul Martin Putora & Rianne Fijten (2025). „The Epistemic Cost of Opacity: How the Use of Artificial Intelligence Undermines the Knowledge of Medical Doctors in High-Stakes Contexts“. In: *Philosophy & Technology*, Bd. 38, Nr. 1, Art. Nr. 5, S. 1–22. doi: 10.1007/s13347-024-00834-9.
- Schubert, Ludwig, Chelsea Voss, Nick Cammarata, Gabriel Goh & Chris Olah (2021). „High-Low Frequency Detectors“. In: *Distill*, Bd. 6, Nr. 1. doi: 10.23915/distill.00024.005.
- Schwalbe, Gesina & Bettina Finzel (2024). „A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts“. In: *Data Mining and Knowledge Discovery*, Bd. 38, S. 3043–3101. doi: 10.1007/s10618-022-00867-8.
- Schwitzgebel, Eric (2024). „Belief“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N. Zalta & Uri Nodelman. Frühlings-Edition. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/spr2024/entries/belief/>.
- Selbst, Andrew D. & Solon Barocas (2018). „The Intuitive Appeal of Explainable Machines“. In: *Fordham Law Review*, Bd. 87, S. 1085–1139. doi: 10.2139/ssrn.3126971.
- Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra (2017). „Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization“. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Venedig, 22.–29. Okt. 2017). Los Alamitos, Washington Tokio: Institute of Electrical and Electronics Engineers (IEEE), S. 618–626. doi: 10.1109/iccv.2017.74.
- Shallue, Christopher J. & Andrew Vanderburg (2018). „Identifying Exoplanets with Deep Learning: A Five-Planet Resonant Chain around Kepler-80 and an Eighth Planet around Kepler-90“. In: *The Astronomical Journal*, Bd. 155, Nr. 2, Art. Nr. 94, S. 1–21. doi: 10.3847/1538-3881/aa9e09.
- Shortliffe, Edward H., Stanton G. Axline, Bruce G. Buchanan, Thomas C. Merigan & Stanley N. Cohen (1973). „An Artificial Intelligence Program to Advise Physicians Regarding

- Antimicrobial Therapy“. In: *Computers and Biomedical Research*, Bd. 6, Nr. 6, S. 544–560. DOI: 10.1016/0010-4809(73)90029-3.
- Simion, Mona & Christoph Kelp (2020). „Conceptual Innovation, Function First“. In: *Nous*, Bd. 54, Nr. 4, S. 985–1002. DOI: 10.1111/nous.12302.
- Skiena, Steven S. (2017). *The Data Science Design Manual*. Texts in Computer Science. Cham: Springer. DOI: 10.1007/978-3-319-55444-0.
- Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh & Himabindu Lakkaraju (2020). „Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods“. In: *Proceedings of the 3rd AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (New York, 7.–9. Feb. 2020). New York: Association for Computing Machinery (ACM), S. 180–186. DOI: 10.1145/3375627.3375830.
- Sleeuw, Stefan (2023). „Understanding for a Purpose“. Diss. Universität Groningen. DOI: 10.33612/diss.568507019.
- Sliwa, Paulina (2015). „IV-Understanding and Knowing“. In: *Proceedings of the Aristotelian Society*, Bd. 115, Nr. 1, Teil 1, S. 57–74. DOI: 10.1111/j.1467-9264.2015.00384.x.
- Smolensky, Paul (1988). „On the Proper Treatment of Connectionism“. In: *Behavioral and Brain Sciences*, Bd. 11, Nr. 1, S. 1–74. DOI: 10.1017/s0140525x00052432.
- Sober, Elliott (1976). „Mental Representations“. In: *Synthese*, Bd. 33, Nr. 2–4, S. 101–148. DOI: 10.1007/bf00484711.
- Sokol, Kacper & Peter Flach (2020). „Explainability Fact Sheets“. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT)* (Barcelona, 27.–30. Jan. 2020). New York: Association for Computing Machinery (ACM), S. 56–67. DOI: 10.1145/3351095.3372870.
- (2022). „Explainability is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence“. arXiv: 2112.14466.
- Søndergaard, Andreas (2023). „Understanding As Usability and Context-Sensitivity to Interests“. In: *Philosophia*, Bd. 51, Nr. 5, S. 2603–2623. DOI: 10.1007/s11406-023-00699-w.
- Speith, Timo (2022). „A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods“. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (Seoul, 21.–24. Juni 2022). New York: Association for Computing Machinery (ACM), S. 2239–2250. DOI: 10.1145/3531146.3534639.
- (2023). „Building Bridges for Better Machines: From Machine Ethics to Machine Explainability and Back“. Lizenziert via CC BY-NC-ND 3.0 DE Namensnennung – Keine kommerzielle Nutzung – Keine Bearbeitungen 3.0 Deutschland, <https://creativecommons.org/licenses/by-nc-nd/3.0/de/>. Diss. Universität des Saarlandes. DOI: 10.22028/D291-40450.
- Speith, Timo, Barnaby Crook, Sara Mann, Astrid Schomäcker & Markus Langer (2024). „Conceptualizing Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach“. In: *Ethics and Information Technology*, Bd. 26, Nr. 2, Art. Nr. 40. Lizenziert via CC BY 4.0 Namensnennung 4.0 International, <https://creativecommons.org/licenses/by/4.0/>, S. 1–15. DOI: 10.1007/s10676-024-09769-3.

- Stanley, Jason (2005). *Knowledge and Practical Interests*. Oxford, New York: Oxford University Press. DOI: 10.1093/0199288038.001.0001.
- Stanley, Jason & Timothy Williamson (2001). „Knowing How“. In: *Journal of Philosophy*, Bd. 98, Nr. 8, S. 411–444. DOI: 10.2307/2678403.
- Stepin, Ilia, Jose M. Alonso, Alejandro Catala & Martin Pereira-Farina (2021). „A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence“. In: *IEEE Access*, Bd. 9, S. 11974–12001. DOI: 10.1109/access.2021.3051315.
- Steup, Matthias, John Turri & Ernest Sosa, Hrsg. (2014). *Contemporary Debates in Epistemology*. Malden, Oxford, Chichester: Blackwell.
- Strawson, Peter F. (1963). „Carnap’s Views on Conceptual Systems Versus Natural Languages in Analytic Philosophy“. In: *The Philosophy of Rudolf Carnap*. Hrsg. von Paul Arthur Schilpp. The Library of Living Philosophers 11. La Salle: Open Court, S. 503–518.
- Strevens, Michael (2013). „No Understanding without Explanation“. In: *Studies in History and Philosophy of Science Part A*, Bd. 44, Nr. 3, S. 510–515. DOI: 10.1016/j.shpsa.2012.12.005.
- (2024). „Grasp and Scientific Understanding: A Recognition Account“. In: *Philosophical Studies*, Bd. 181, Nr. 4, S. 741–762. DOI: 10.1007/s11098-024-02121-x.
- Stuart, Michael T. (2016). „Taming Theory with Thought Experiments: Understanding and Scientific Progress“. In: *Studies in History and Philosophy of Science Part A*, Bd. 58, S. 24–33. DOI: 10.1016/j.shpsa.2016.04.002.
- (2025). „A New account of Pragmatic Understanding, applied to the Case of AI-Assisted Science“. In: *Philosophical Studies*. DOI: 10.1007/s11098-025-02336-6.
- Sullivan, Emily (2018). „Understanding: Not Know-How“. In: *Philosophical Studies*, Bd. 175, Nr. 1, S. 221–240. DOI: 10.1007/s11098-017-0863-z.
- (2022). „Understanding from Machine Learning Models“. In: *The British Journal for the Philosophy of Science*, Bd. 73, Nr. 1, S. 109–133. DOI: 10.1093/bjps/axz035.
- (2023). „How Values Shape the Machine Learning Opacity Problem“. In: *Scientific Understanding and Representation: Modeling in the Physical Sciences*. Hrsg. von Insa Lawler, Kareem Khalifa & Elay Shech. Routledge Studies in the Philosophy of Mathematics and Physics. New York, Oxon: Routledge, S. 306–322. DOI: 10.4324/9781003202905-27.
- Sun, Ron (2014). „Connectionism and Neural Networks“. In: *The Cambridge Handbook of Artificial Intelligence*. Hrsg. von Keith Frankish & William M. Ramsey. Cambridge: Cambridge University Press, S. 108–127. DOI: 10.1017/cbo9781139046855.008.
- Suresh, Harini, Steven R. Gomez, Kevin K. Nam & Arvind Satyanarayan (2021). „Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs“. In: *Proceedings of the 2021 Conference on Human Factors in Computing Systems (CHI)* (Yokohama, 8.–13. Mai 2021). New York: Association for Computing Machinery (ACM), Art. Nr. 74, S. 1–16. DOI: 10.1145/3411764.3445088.
- Swartout, William R. & Johanna D. Moore (1993). „Explanation in Second Generation Expert Systems“. In: *Second Generation Expert Systems*. Hrsg. von Jean-Marc David, Jean-Paul

- Krivine & Reid Simmons. Berlin, Heidelberg: Springer, S. 543–585. DOI: 10.1007/978-3-642-77927-5_24.
- Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow & Rob Fergus (2014). „Intriguing Properties of Neural Networks“. arXiv: 1312.6199.
- Tensorflow (23. Juli 2026). *Models: Models and examples built with TensorFlow*. Lizenziert via Apache-Lizenz 2.0. GitHub. URL: <https://github.com/tensorflow/models> (besucht am 05.08.2026).
- Thomasson, Amie L. (2024). „How Should We Think about Linguistic Function?“ In: *Inquiry*, Bd. 67, Nr. 3, S. 840–871. DOI: 10.1080/0020174x.2022.2074886.
- Thórisson, Kristinn R., David Kremelberg, Bas R. Steunebrink & Eric Nivel (2016). „About Understanding“. In: *Proceedings of the 9th International Conference on Artificial General Intelligence (AGI)* (New York, 16.–19. Juli 2016). Cham: Springer, S. 106–117. DOI: 10.1007/978-3-319-41649-6_11.
- Tomsett, Richard, Dave Braines, Dan Harborne, Alun Preece & Supriyo Chakraborty (2018). „Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems“. In: *Proceedings of the 3rd Annual Workshop on Human Interpretability in Machine Learning (WHI)* (Stockholm, 14. Juli 2018). Hrsg. von Been Kim, Kush R. Varshney & Adrian Weller. 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), S. 8–14. arXiv: 1806.07552.
- Trout, John Dewain (2002). „Scientific Explanation and the Sense of Understanding“. In: *Philosophy of Science*, Bd. 69, Nr. 2, S. 212–233. DOI: 10.1086/341050.
- (2007). „The Psychology of Scientific Explanation“. In: *Philosophy Compass*, Bd. 2, Nr. 3, S. 564–591. DOI: 10.1111/j.1747-9991.2007.00081.x.
- van Camp, Wesley (2014). „Explaining Understanding (or Understanding Explanation)“. In: *European Journal for Philosophy of Science*, Bd. 4, Nr. 1, S. 95–114. DOI: 10.1007/s13194-013-0077-y.
- Van Cleve, James (2014). „Why Coherence Is Not Enough: A Defense of Moderate Foundationalism“. In: *Contemporary Debates in Epistemology*. Hrsg. von Matthias Steup, John Turri & Ernest Sosa. Malden, Oxford, Chichester: Blackwell, S. 255–267.
- van Fraassen, Bas C. (1980). *The Scientific Image*. Oxford, New York: Oxford University Press. DOI: 10.1093/0198244274.001.0001.
- van Otterlo, Martijn (2013). „A Machine Learning View on Profiling“. In: *Privacy, Due Process and the Computational Turn. The Philosophy of Law Meets the Philosophy of Technology*. Hrsg. von Mireille Hildebrandt & Katja de Vries. London: Routledge, S. 41–64.
- van Riel, Raphael (2015). „The Content of Model-Based Information“. In: *Synthese*, Bd. 192, Nr. 12, S. 3839–3858. DOI: 10.1007/s11229-015-0728-y.
- Vanello, Daniel (2025). „Moral Understanding, Affect, and the Imagination“. In: *Inquiry*, Bd. 68, Nr. 2, S. 183–208. DOI: 10.1080/0020174x.2022.2077431.
- von Luxburg, Ulrike, Robert C. Williamson & Isabelle Guyon (2012). „Clustering: Science or Art?“ In: *Proceedings of the ICML Workshop on Unsupervised and Transfer Learning*

- (Bellevue, 2. Juli 2011). Hrsg. von Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor & Daniel Silver. Bd. 27. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research (PMLR), S. 65–79. URL: <https://proceedings.mlr.press/v27/luxburg12a.html>.
- Wachter, Sandra, Brent Daniel Mittelstadt & Chris Russell (2018). „Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR“. In: *Harvard Journal of Law & Technology*, Bd. 31, Nr. 2, S. 841–887. DOI: 10.2139/ssrn.3063289.
- Wagenknecht, Susann (2014). „Opaque and Translucent Epistemic Dependence in Collaborative Scientific Practice“. In: *Episteme*, Bd. 11, Nr. 4, S. 475–492. DOI: 10.1017/epi.2014.25.
- Wen, David, Saad M. Khan, Antonio Ji Xu, Hussein Ibrahim, Luke Smith, Jose Caballero, Luis Zepeda, Carlos de Blas Perez, Alastair K. Denniston, Xiaoxuan Liu & Rubeta N. Matin (2021). „Characteristics of Publicly Available Skin Cancer Image Datasets: A Systematic Review“. In: *The Lancet Digital Health*, Bd. 4, Nr. 1, S. e64–e74. DOI: 10.1016/s2589-7500(21)00252-1.
- White, Adam, Kwun Ho Ngan, James Phelan, Kevin Ryan, Saman Sadeghi Afgeh, Constantino Carlos Reyes-Aldasoro & Artur d’Avila Garcez (2023). „Contrastive Counterfactual Visual Explanations with Overdetermination“. In: *Machine Learning*, Bd. 112, Nr. 9, S. 3497–3525. DOI: 10.1007/s10994-023-06333-w.
- Wilkenfeld, Daniel A. (2013). „Understanding As Representation Manipulability“. In: *Synthese*, Bd. 190, Nr. 6, S. 997–1016. DOI: 10.1007/s11229-011-0055-x.
- (2015). „The Contextual Quasi-Factivity of Objectual Understanding“. Manuskript.
 - (2017a). „MUDdy Understanding“. In: *Synthese*, Bd. 194, Nr. 4, S. 1273–1293. DOI: 10.1007/s11229-015-0992-x.
 - (2017b). „Understanding without Believing“. In: *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Hrsg. von Stephen R. Grimm, Christoph Baumberger & Sabine Ammon. New York, Oxon: Routledge, S. 318–333.
- Williams, Joseph J. & Tania Lombrozo (2010). „The Role of Explanation in Discovery and Generalization: Evidence from Category Learning“. In: *Cognitive Science*, Bd. 34, Nr. 5, S. 776–806. DOI: 10.1111/j.1551-6709.2010.01113.x.
- Winkler, Julia K., Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz & Holger A. Haenssle (2019). „Association between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition“. In: *JAMA Dermatology*, Bd. 155, Nr. 10, S. 1135–1141. DOI: 10.1001/jamadermatol.2019.1735.
- Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford, New York: Oxford University Press. DOI: 10.1093/0195155270.001.0001.
- (2014). „Explanation“. In: *The Routledge Companion to Philosophy of Science*. Hrsg. von Martin Curd & Stathis Psillos. 2. Aufl. Oxon, New York: Routledge, S. 203–213.
- Wu, Xiaolin & Xi Zhang (2016). „Responses to Critiques on Machine Learning of Criminality Perceptions (addendum of Arxiv:1611.04135)“. arXiv: 1611.04135.

- Ylikoski, Petri K. (2009). „The Illusion of Depth of Understanding in Science“. In: *Scientific Understanding: Philosophical Perspectives*. Hrsg. von Henk de Regt, Sabina Leonelli & Kai Eigner. Pittsburgh: University of Pittsburgh Press, S. 100–119.
- Ylikoski, Petri K. & Jaakko Kuorikoski (2010). „Dissecting Explanatory Power“. In: *Philosophical Studies*, Bd. 148, Nr. 2, S. 201–219. doi: 10.1007/s11098-008-9324-z.
- Yu, Rulei & Lei Shi (2018). „A User-Based Taxonomy for Deep Learning Visualization“. In: *Visual Informatics*, Bd. 2, Nr. 3, S. 147–154. doi: 10.1016/j.visinf.2018.09.001.
- Zagzebski, Linda Trinkaus (1996). *Virtues of the Mind. An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139174763.
- (2001). „Recovering Understanding“. In: *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*. Hrsg. von Matthias Steup. New York: Oxford University Press, S. 235–252. doi: 10.1093/0195128923.003.0015.
- (2009). *On Epistemology*. Wadsworth Philosophical Topics. Belmont: Wadsworth, Cengage Learning.
- (2019). „Toward a Theory of Understanding“. In: *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*. Hrsg. von Stephen R. Grimm. New York: Oxford University Press, S. 123–136. doi: 10.1093/oso/9780190860974.003.0007.
- Zhou, Jianlong, Amir H. Gandomi, Fang Chen & Andreas Holzinger (2021). „Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics“. In: *Electronics*, Bd. 10, Nr. 5, Art. Nr. 593, S. 1–19. doi: 10.3390/electronics10050593.
- Zhou, Zhenglong & Chaz Firestone (2019). „Humans Can Decipher Adversarial Images“. In: *Nature Communications*, Bd. 10, Art. Nr. 1334, S. 1–9. doi: 10.1038/s41467-019-08931-6.