

Understanding and the Black Box Problem in Artificial Intelligence

Eine Dissertation

Der Fakultät
für Kulturwissenschaft
der Universität Bayreuth
zur Erlangung des akademischen Grades
Doktor der Philosophie

Eingereicht von: Barnaby William Crook, M.A.
am 28.11.2024

Fakultät: Kulturwissenschaftlichen Fakultät der Universität Bayreuth

Dekanin: Prof. Dr. Paula Schrode

Betreuerin: Prof. Dr. Lena Kästner

Gutachterin: Prof. Dr. Lena Kästner

Gutachterin: Prof. Dr. Eva Schmidt

Tag der Annahme der Arbeit: 5. Februar 2025

Abstract

(Deutsch)

Künstliche Intelligenz (KI)-Systeme, die mit Verfahren des maschinellen Lernens entwickelt werden, finden breite Anwendung in Industrie, Wissenschaft und Gesellschaft. Die Intransparenz dieser Systeme droht jedoch, die Erfüllung grundlegender Desiderate wie Sicherheit, Vertrauen und Fairness zu untergraben. Dieses Problem ist als das „Black-Box-Problem“ (BBP) bekannt. In dieser Dissertation entwickle ich ein neuartiges konzeptionelles Framework zur Analyse des BBP: Durch fünf miteinander verbundene wissenschaftliche Artikel behandle ich drei zentrale Forschungsfragen zu seiner Natur und unseren Aussichten, es zu lösen.

Zunächst betrachte ich den zentralen epistemischen Begriff des Verstehens und argumentiere, dass eine auf Fähigkeiten basierende Auffassung von Verstehen den Anforderungen der Analyse des BBP am besten gerecht wird. Das Konzeptualisieren von Verstehen als Fähigkeiten, so behaupte ich, ermöglicht eine angemessene Balance zwischen theoretischer Fundierung und praktischer Anwendbarkeit und kann die interdisziplinäre Zusammenarbeit im Bereich erklärbarer KI (XAI) fördern. Zweitens befasse ich mich mit dem Potenzial von XAI, einem Forschungsfeld, das darauf abzielt, KI-Systeme für menschliche Akteure nachvollziehbar zu machen. Mein Schwerpunkt liegt auf der *Mechanistic Interpretability*, einem Ansatz, der funktionale Zerlegung nutzt, um die interne Struktur von KI-Systemen zu beschreiben. Mechanistic Interpretability kann detaillierte Erklärungen liefern, erfordert jedoch erheblichen Zeit- und Arbeitsaufwand. Darüber hinaus stelle ich ein Framework zur Bewertung von Erklärbarkeitsansätzen vor und zeige, wie man es nutzen kann, um den vermeintlichen Konflikt zwischen Leistung und Erklärbarkeit effektiv zu analysieren. Drittens untersuche ich die Lösbarkeit des BBP in spezifischen Kontexten. In der Neurowissenschaft argumentiere ich gegen Kritiker, dass künstliche neuronale Netzwerkmodelle in Verbindung mit XAI unser Verständnis des Gehirns fördern können. Bezüglich der Psychiatrie hingegen argumentiere ich, dass datenbasierte Ansätze selbst dann, wenn die Intransparenz von KI-Systemen überwunden wird, wahrscheinlich keine transformativen Verbesserungen der Behandlungsergebnisse liefern werden.

Insgesamt zeigt diese Dissertation, wie eine auf Fähigkeiten basierende Perspektive des Verstehens und eine fundierte Bewertung des Potenzials von XAI kombiniert werden können, um neue Erkenntnisse über das BBP zu gewinnen und dessen Lösung in unterschiedlichen Kontexten zu fördern.

(English)

Artificial intelligence (AI) systems developed with automated machine learning procedures are used throughout industry, science, and society. However, the opacity of these systems threatens to undermine the satisfaction of critical desiderata including safety, trust, and fairness. This worry is known as the *black box problem* (BBP). In this dissertation, I develop a novel conceptual framework for analysing the BBP and, through five interrelated papers, address three key research questions about its nature and our prospects for solving it.

First, I consider the central epistemic construct of *understanding*, arguing that an abilities-based view is best-suited to the requirements of BBP analysis. Conceptualising understanding as abilities, I claim, appropriately balances theoretical justification with practical utility and can help coordinate interdisciplinary efforts in explainable AI (XAI). Second, I address the potential of XAI, a research field which attempts to render AI systems comprehensible to human stakeholders. I focus on *mechanistic interpretability*, an approach which employs functional decomposition to characterise the internal structure of AI systems. Mechanistic interpretability can deliver detailed explanations, but at significant costs in time and labour. Additionally, I present a framework for evaluating explainability approaches, using it to show how performance-explainability trade-offs can be effectively analysed. Third, I address the solubility of the BBP in particular contexts. In neuroscience, I argue against critics that artificial neural network models, in conjunction with XAI, can enhance our understanding of the brain. In psychiatry, I argue that even if the opacity of AI systems is overcome, data-driven approaches are unlikely to deliver transformative improvements to patient outcomes.

Overall, this dissertation demonstrates how an abilities-based view of understanding and a considered assessment of XAI's potential can be combined to yield new insights into the BBP and inform its resolution across diverse contexts.

Declaration of Authorship

(Deutsch)

Ich versichere hiermit an Eides Statt, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Darüber hinaus versichere ich, dass ich weder bisher Hilfe von gewerblichen Promotionsberatern bzw. -vermittlern in Anspruch genommen habe noch künftig in Anspruch nehmen werde. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht.

(English)

I hereby do solemnly affirm that I have produced this work without the unauthorised help of third parties and without the use of sources or aids other than those cited; the ideas stemming directly or indirectly from external sources are marked as such. I do further declare that I have not enlisted the help of professional thesis consultants or coaches, neither will I use them in future. This work has not previously been submitted to an examining authority in Germany or abroad, neither in the same nor a similar form, and has also not been published.

27.11.2024

Barnaby William Crook

Curriculum Vitae

Barnaby Crook

Universität Bayreuth, Institut für Philosophie, 95440 Bayreuth, Germany

Barnaby.Crook@uni-bayreuth.de, +447757551492

Education History

- 04.2022 – Present **PhD Candidate in Philosophy of Artificial Intelligence** – Bayreuth University, Germany
- 06.2021 – 04.2022 **PhD Candidate in Philosophy of Artificial Intelligence** – Saarland University, Germany
- 09.2018 – 06.2021 **MA in Philosophy and Neuroscience** – Berlin School of Mind and Brain, Humboldt University, Berlin, Germany
- 09.2009 – 06.2012 **BA in Philosophy** – University of Leeds, UK

Scientific Papers

- Risks Deriving from the Agential Profiles of Modern AI Systems. Forthcoming in *Philosophy of Artificial Intelligence: The State of the Art*.
- Normativity and the Intelligent Matter Framework. Forthcoming in *Artificial Intelligence and Intelligent Matter*.
- Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience (with Lena Kästner). Forthcoming in *Philosophy of Science for Machine Learning: Core Issues, New Perspectives*.
- Explaining AI Through Mechanistic Interpretability (with Lena Kästner). *European Journal for the Philosophy of Science*, 2024.
- Conceptualizing Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach (with Timo Speith, Sara Mann, Astrid Schomäcker, and Markus Langer). *Ethics and Information Technology*, 2024.
- Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (with Maximilian Schlüter and Timo Speith). *International Requirements Engineering Conference Workshops*, 2023.
- Understanding and Addressing Sources of Opacity in Computer Systems (with Sara Mann, Lena Kästner, Astrid Schomäcker, and Timo Speith). *International Requirements Engineering Conference Workshops*, 2023.
- Understanding as a Bottleneck for the Data-Driven Approach to Psychiatric Science. *Philosophy and the Mind Sciences*, 4, 2023.

Other Published Writing

Review of “The Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together”, Luis Pessoa. *H-Net Reviews*, 2023.

Iterative Tweaking and the Interpretation of Feature Visualisation. *Explainable Intelligent Systems Blog*, 2021.

Mastering Motor Movements: How Caching and Circuits Support Motor Learning. *Charite Neuroscience Magazine*, 2021.

Lessons From the Pandemic: What SARS-CoV-2 can teach us about making decisions under uncertainty. *Charite Neuroscience Magazine*, 2020.

Academic Talks, Workshops, and Posters

- 15.11.2024 The Structure of Arguments about (Artificial) Intelligence: A Lakatosian Analysis (poster presentation). *Philosophy of Science Association Conference 2024*, New Orleans.
- 09.02.2024 Risks Deriving from the Agential Profiles of Modern AI Systems (poster presentation). *Bayreuth Lab for Digital Sciences Day*, Bayreuth.
- 15.12.2023 Risks Deriving from the Agential Profiles of Modern AI Systems (poster presentation). *Philosophy of Artificial Intelligence*, Erlangen.
- 22.09.2023 Symposium: Explainable AI in Scientific Research (with Lena Kästner). *European Philosophy of Science Association 2023*, Belgrade.
- 04.09.2023 Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence. *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, Hannover.
- 14.07.2023 The Structure of Arguments About Intelligence. *Artificial Intelligence and Intelligent Matter: An Interdisciplinary Perspective*, Münster.
- 04.07.2023 Explaining AI Through the Scientific Perspective. Invited Talk, Ulm University.
- 14.02.2023 What is ChatGPT and How does it Work? *Explainer for the Faculty of Humanities and Social Sciences*, Bayreuth.
- 18.11.2022 A History of AI in 10 Concepts. *Digital Sciences Lecture Series*, Bayreuth.
- 19.10.2022 Mary Hesse, Machine Learning, and the Learning Machine. *Philosophy Meets Machine Learning Graduate Workshop*, Tübingen.
- 15.09.2022 Discovering Emergent Structure in AI Systems (with Lena Kästner). *GAP*, Berlin.
- 16.08.2022 The Compact Core – Emergent Structure Distinction in Artificial and Biological Neural Networks. *GWP*, Berlin.
- 03.06.2022 Elements of Intelligence in Artificial Systems. *Global Classroom Special Friday Workshop*, Saarland.
- 28.03.2022 Agency and the Environment in Artificial Systems (poster presentation). *Situated Cognition: Applications, Prospects and Limits*, Bochum.

Teaching

Summer 2024	Seminar for General Studies course on the Bayreuth Lecture: “Explainable AI with Large Language Models: opportunities, challenges and risks”. University of Bayreuth. Level: undergraduates.
Summer 2024	“Philosophy of Artificial Intelligence” (with Lena Kästner). University of Bayreuth. Level: undergraduates.
Winter 2023/24	“Perspectives on Intelligence: From Philosophy to AI”. University of Bayreuth. Level: undergraduates and MA.
Summer 2023	“Machine Learning in Neuroscience” (with Lena Kästner). University of Bayreuth. Level: undergraduates and MA.
Winter 2021/22	“Artificial Intelligence – History, Core Concepts, and Recent Debates” (with Lena Kästner and Astrid Schomäcker. University of Saarland. Level: MA.

Employment History

06.2021 – present	Research Associate – Explainable Intelligent Systems, Bayreuth, Germany
10.2020 – 06.2021	Grant Writer – Open Energy Labs, London, UK
09.2017 – 06.2018	English Teacher – American Eagle Institute, Zhubei, Taiwan
09.2015 – 08.2017	English Teacher – Korrnell Academy, Zhubei, Taiwan
05.2015 – 09.2015	English Teacher – International House, Bristol, UK
11.2012 – 01.2015	English Teacher – English Hub, Geoje, South Korea

Acknowledgements

The work which makes up this dissertation was started in 2021. Although I began my PhD in June, I was inspired to write the paper which constitutes the 6th chapter of this work by the Minds, Models, and Mechanisms conference which took place completely online (due to Covid) in April of that year. Over the course of the PhD I have received guidance and support from many fantastic people. I won't be able to thank everybody, but I want to mention some of those who've been particularly pivotal in helping me along this journey.

First, I want to thank my supervisor, Lena Kästner. I applied for a PhD position with Lena because I had read and enjoyed her work. Having the chance to collaborate on several projects, two of which are chapters in this dissertation, was extremely rewarding. Lena has been encouraging, kind, and supportive throughout the process and always made time for me when I needed it. I have learned a lot from her along the way about writing, thinking, and life. Thank you.

Next, I want to thank everyone else who collaborated with me on works included in this dissertation. The “abilities paper”, as we referred to CU during the writing process, was originally intended to be a quick group project. Instead, it took three years and many long and arduous meetings to complete. To Timo Speith, Sara Mann, Astrid Schomäcker, and Markus Länger, thank you for bearing with it. Fortunately, the writing process for RPET was much more straightforward. To Timo Speith (again!) and Maximilian Schlüter, thank you for being part of it.

I also want to thank my other colleagues in the EIS project who have played an important role in shaping my thinking about the black box problem and plenty of other topics besides. To Andreas Sesing, Kevin Baum, Eva Schmidt, Thorsten Helfer, Felix Kares, Daniel Oster, Holger Hermanns, Sara Sterz, Ulla Wessels, Tim Hunsicker and everybody else who was part of the EIS team in the last four years, thank you for making it an enriching and fun thing to be part of.

There are many other people in the world of science and academia who have helped me develop, acted as sounding boards for my ideas, or just tickled my brain with fascinating and enjoyable chats over the past 4 years. I cannot possibly hope to thank all of these people individually, but special thanks go to the members of the GRIST reading group

(Jonas Hartmann, Namid Stillman, Joshua Martin, Marika Constant, and Giles Harper-Donelly) which was always a stimulating distraction and a great source of interdisciplinary perspectives on the science of intelligent systems. Several arguments in this dissertation were first proposed there. Thanks also to the anonymous reviewers whose careful reading and probing questions strengthened the argumentation contained in this dissertation immensely.

During this PhD I have spent significant chunks of time in Berlin, Bristol, Saarbrücken, Nürnberg, and Toronto. Every step of the way, amazing friends have been there to celebrate with me when the going is good and commiserate with me when there are bumps in the road. Special thanks and unending friendship go to Sveta Sokolova, James Dunn-Williamson, Grga Grgurevic, Eóin Dooley, Martin Gell, Chiara Caporuscio, Marika Constant, Ryan Lozon, KayLee Flower, Susana Pérez Blázquez, Richard Marsh, Emre Fatih, Leo Auri, Joe Jenkinson, Logan Hadley, Peter Gibbs, Goerge Jenkinson, Callum McClellan, Sean Lemoir, Sam McClellan, James Finn, Will Floutier, Delilah Rooney, Richard Eastick, Leah Morgan, Dan Cech, Olive Baldwin, Catherine Floutier, Kiarash Darafshi, Tom Burns, Samson Sahmland Bowling, Lila Beesley, Yaman Abdin, David Rotele, Joanna Calderón, Agnes Obermann, Emilio Meza, Luisa Maria Arevalo, Chris Klein, and the many others who I'm forgetting right now.

I also want to thank my brilliant family: Chris Crook, Jenny Mulvaney, Alice Crook, Harry Kauffman, Nina Crook, Sid Crook, Livi Ogilvie, Henry Bell, Maili Bell, Frankie Bell and Soren Bell. And my partner's brilliant family: Karl Waclawek, Diane Waclawek, Margaret Boehm, Greg Waclawek, and Katharina Waclawek.

Finally, extremely special thanks to my wonderful, loving, supportive partner Theresa Waclawek. I couldn't have made it through without you.

Funding

Work on this paper was funded by the Volkswagen Foundation grant "Explainable Intelligent Systems" (EIS) (grant numbers 9B 830 and 98 509). The Volkswagen Foundation had no role in preparation, review, or approval of the manuscript. The author declares no other financial interests.

Contents

<i>List of Tables and Figures</i>	<i>xii</i>
<i>Abbreviations</i>	<i>xiii</i>
<i>Overview of Research Questions and Contributions</i>	<i>xiv</i>
Chapter 1: Introduction	1
1.1 Motivation and Research Questions.....	1
1.2 Scientific and Philosophical Background	4
1.3. The Black-Box Problem	16
1.4 My Contributions.....	28
1.5 Summary	43
1.6 References	44
Chapter 2: Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-based Approach	63
2.0 Abstract.....	63
2.1 Introduction	64
2.2 Criteria for a Useful Conceptualisation of Understanding in XAI	66
2.3 Abilities Are Motivated by Research Across Disciplines	70
2.4. Abilities Are Versatile.....	72
2.5 Abilities Are Operationalisable	79
2.6 Abilities Can Be Linked to Desiderata	81
2.7 Limitations and Objections	85
2.8 Conclusion	88
2.9 References	89
Chapter 3: Explaining AI Through Mechanistic Interpretability	97
3.0 Abstract.....	97
3.1 Introduction	97

3.2 Contemporary XAI: The Divide-and-Conquer Strategy	100
3.3 Augmenting Contemporary XAI Through Mechanistic Interpretability	105
3.4 Six Worries ... and Responses	116
3.5 Conclusion	120
3.6 References	121
 Chapter 4: Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)	
4.0 Abstract.....	131
4.1 Introduction	131
4.2 The Performance-Explainability Trade-Off	133
4.3 Contra Rudin's Critique	135
4.4 Re-Imagining the Performance-Explainability Trade-Off	143
4.5 Conclusion	147
4.6 References	147
 Chapter 5: Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience.....	
5.0 Abstract.....	155
5.1 Introduction	156
5.2 Trade-Off Worries	158
5.3 Understanding the Brain with Deep Learning	168
5.4 Conclusion	177
5.5 References	178
 Chapter 6: Understanding as a Bottleneck for the Data-Driven Approach to Psychiatric Science.....	
6.0 Abstract.....	185
6.1 Introduction	186
6.2 The Data-driven Approach.....	188

6.3 The Bottleneck of Understanding.....	199
6.4 The Data-driven Approach and the Bottleneck of Understanding.....	204
6.5 Responses to Objections	211
6.6 Conclusion	214
6.7 References	215
Chapter 7: Conclusion.....	225
7.1 Addressing the Wickedness of the Black Box Problem.....	225
7.2 Understanding, Explainability Approaches, and Desiderata Satisfaction: Addressing the Research Questions	227
7.3 Looking Forward: Open Questions and Future Work	232
7.4 References	235
Appendix A: Clarification of Specific Contributions.....	241
Full Bibliography.....	243

List of Tables and Figures

Tables

Table 1: Overview of contributions.....	xv
Table 2: Overview of explainability approaches.....	22

Figures

Figure 1: Research Questions in Context.....	xiv
Figure 2: A Model of the Black Box Problem	26
Figure 3: The Research Focus of CU in Context.....	31
Figure 4: The Research Focus of XAI-MI in Context	34
Figure 5: The Research Focus of RPET in Context	36
Figure 6: The Research Focus of DFB in Context	39
Figure 7: The Research Focus of UB in Context	42
Figure 8: A Simplified Model of Explainability.....	69
Figure 9: A Loose Hierarchy of Clusters of Abilities.....	73
Figure 10: Multi-level Mechanistic Explanation	108
Figure 11: An Idealised Depiction of PET+	144
Figure 12: Factors Influencing PET+	145

Abbreviations

General

AI = Artificial Intelligence

ML = Machine Learning

DL = Deep Learning

ANN = Artificial Neural Network

DNN = Deep Neural Network

XAI = Explainable Artificial Intelligence

BBP = Black Box Problem

PET = Performance Explainability Trade-Off

MI = Mechanistic Interpretability

Paper Titles

CU = *Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach*

XAI-MI = *Explaining AI Through Mechanistic Interpretability*

RPET = *Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)*

DFB = *Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience*

UB = *Understanding as a Bottleneck for the Data-driven Approach to Psychiatric Science*

Overview of Research Questions and Contributions

LIST OF RESEARCH QUESTIONS

RQ1 = How should *understanding* be construed in the context of the BBP?

RQ2 = How effective are extant explainability approaches?

RQ3 = To what extent can specific manifestations of the BBP be solved?

FIGURE 1: RESEARCH QUESTIONS IN CONTEXT

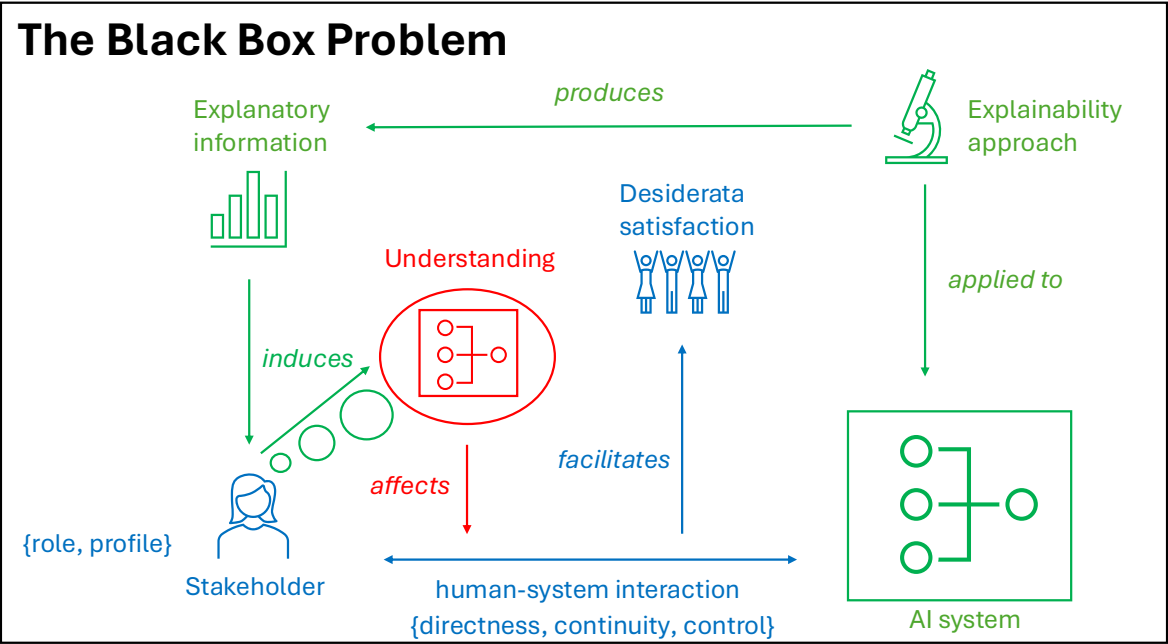


Figure 1 depicts my conceptual model of the black box problem. The colour coding highlights the focus of each research question. The figure is repeated with further explanation and discussion in Chapter 1.3.2.

TABLE 1: OVERVIEW OF CONTRIBUTIONS

(see overleaf)

Title	Details	Description
<i>Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach</i> (CU)	Research Question: 1 Co-Authors: Timo Speith, Sara Mann, Astrid Schomäcker, and Markus Langer Publication Venue: Ethics and Information Technology	We present and defend an approach to understanding that identifies it with the possession of particular <i>abilities</i> (recognising, assessing, predicting, intervening, explaining, designing). We argue that this approach balances theoretical motivation and practical applicability, and is particularly well-suited to XAI contexts.
<i>Explaining AI Through Mechanistic Interpretability</i> (XAI-MI)	Research Question: 2 Co-Author: Lena Kästner Publication Venue: European Journal for the Philosophy of Science	We provide a philosophical analysis of the <i>mechanistic interpretability</i> (MI) approach to XAI. We construe MI as a <i>pattern discovery</i> process employing strategies like functional decomposition to yield detailed mechanistic explanations. MI affords understanding and control of target systems, but at significant costs in terms of time and labour.
<i>Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)</i> (RPET)	Research Question: 2 Co-Authors: Maximilian Schlüter and Timo Speith Publication Venue: 2023 International Requirements Engineering Conference	We critically analyse the hypothesised trade-off between the <i>performance</i> of AI models and their <i>explainability</i> . We argue that a careful analysis of the purported trade-off must incorporate considerations of resources availability, domain complexity, and the risks involved with deploying an AI system in the particular application context.
<i>Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience</i> (DFB)	Research Question: 3 Co-Author: Lena Kästner Publication Venue: Philosophy of Science for Machine Learning: Core Issues and New Perspectives	We critique claims that the use of deep learning in neuroscience entails the sacrifice of scientific understanding. Instead, we argue, developing artificial neural network-based models of cognitive and perceptual capacities offers a novel epistemic perspective that complements traditional approaches and affords unique insights.
<i>Understanding as a Bottleneck for the Data-driven Approach to Psychiatric Science</i> (UB)	Research Question: 3 Co-Authors: None Publication Venue: Philosophy and the Mind Sciences	I critique claims that a data-driven approach to clinical decision-making will prove <i>transformative</i> for psychiatry. I argue that an inability to target patient understanding, construed as possession of adaptive mental constructs, limits the potential of the data-driven approach to improve patient outcomes.

Chapter 1:

Introduction

“An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside”

Alan Turing, *Computing Machinery and Intelligence*, 1950, p.459

1.1 Motivation and Research Questions

In recent years, the use of artificial intelligence (AI) has become increasingly prevalent in science and society (Agrawal et al., 2024). While AI is a broad field, the current wave of progress has been driven primarily by *deep learning* (DL) (Goodfellow et al., 2016; LeCun et al., 2015). In DL, artificial neural networks (ANNs) are trained with large volumes of data to approximate target functions. The DL training procedure is automated, leveraging general-purpose machine learning algorithms to iteratively tweak parameters until they perform well according to a quantified objective. Numerous successes across domains demonstrate that this data-driven approach to AI is an effective means of developing powerful systems (e.g., T. Brown et al., 2020; Jumper et al., 2021; Krizhevsky et al., 2012; Silver et al., 2017). However, these achievements come with a catch. The training procedure produces enormously complex models which are, *prima facie*, opaque (Burrell, 2016). In particular, *how* trained ANNs implement their behaviour is often a mystery, even to their developers (Creel, 2020). For this reason, DL models are referred to as *black boxes* (Durán & Jongsma, 2021; Rudin, 2019). Since we would frequently like to assess the reasoning underlying model behaviour or identify the data patterns they exploit, this poses a problem (Baum et al., 2022; Boge, 2022). The challenge arising from our impoverished epistemic grasp of modern AI systems is known as the *black box problem* (BBP) (Zednik, 2021).¹ In response to the BBP, the field of explainable AI (XAI) attempts to render opaque AI systems explainable and understandable to human stakeholders (Belle & Papantonis, 2021; Guidotti et al., 2021; Langer et al., 2021).

¹ I use the terms *ANN* and *DL model* interchangeably. The term *AI system* is broader (see Chapter 2.1 for more detail).

Chapter 1: Introduction

At first pass the BBP may seem like a technical problem to be addressed by computer scientists. In reality, however, the BBP encompasses a sprawling and interrelated set of social, technical, scientific, and philosophical problems (Langer et al., 2021). For one thing, XAI pursues epistemic goals (explainability and understanding), but how these goals should be interpreted, operationalised, and evaluated is unclear and controversial (Erasmus et al., 2021; Fleisher, 2022; Krishnan, 2020; Vilone & Longo, 2021). For another, the explanatory targets of XAI can range from single model predictions (Ribeiro et al., 2016), to the functional structure of complex trained models (Olah et al., 2020), to the wider sociotechnical systems in which models are deployed (e.g., the legal system) (Vredenburg, 2022). For yet another, DL is a highly general methodology applicable across countless domains. As such, the stakeholders interacting with DL-based AI systems vary enormously with respect to their background knowledge, societal obligations, and epistemic desiderata (Imrie et al., 2023; Langer et al., 2021; Tomsett et al., 2018). In light of these considerations, the BBP can be disambiguated into an enormous number of distinct problems depending on the stakeholder for whom the relevant epistemic state ought to be induced, the goal for which that epistemic state is hypothesised to be instrumental, the system aspect which constitutes the explanatory target, and the degree of explainability or understanding required. Adequate consideration of these factors is critical to the practical success of XAI (Langer et al., 2021). Thus, interdisciplinary work is required to align, organise, and direct research in the field.

I take the considerations above to show that “the” BBP is really a variegated cluster of problems (see also Zednik, 2021). To distinguish such problems from the BBP writ large I will call them *manifestations* of the BBP in what follows (see Chapter 1.3.1). The various manifestations of the BBP are united by a common cause, namely, the opacity of AI systems (Boge, 2022; Facchini & Termine, 2022; Mann et al., 2023; Zednik, 2021). While the heterogeneity of the BBP renders general solutions infeasible, (Belle & Papantonis, 2021; Nyrup & Robinson, 2022), its manifestations share conceptual terms (e.g., explanation, desideratum, stakeholder) and can be analysed in a single conceptual framework. In this introduction, I present my own framework for the BBP and use it to organise the contributions of my dissertation. The framework consists of construals of

the conceptual components of the BBP (Chapter 1.3.1) and a novel model of how they interact (Chapter 1.3.2). Homing in on particular facets of the conceptual framework (i.e., specific components and/or their relations) reveals numerous questions which demand philosophical consideration. My dissertation focuses on three such questions. First, how should the core conceptual terms be understood? For reasons elaborated in Chapter 1.2.2, I focus specifically on what it means to *understand* an AI system. This leads to **Research Question (RQ) 1: How should *understanding* be construed in the context of the BBP?** In my view, a pragmatic analysis tailored to the specificities of the BBP is required. Such an analysis ought to provide a construal flexible enough to be applied across widely diverse contexts (viz., BBP manifestations) and concrete enough to guide technical and empirical research (see Chapter 1.4.1 for details). Second, to what extent *can* XAI ameliorate the opacity of these systems? This is the subject of **RQ2: How effective are extant explainability approaches?** Answering this question requires characterising the XAI landscape in a way that accurately captures the strengths and weaknesses of different approaches (see Chapter 1.4.2). Third, moving from the general to the specific, how does the opacity of AI systems impede the desiderata satisfaction of specific stakeholders and does XAI offer promising solutions? In other words, **RQ3: To what extent can specific manifestations of the BBP be solved?** Here, domain-specific analyses are required. Such analyses ought to focus on the particularities of the domain in question to fix details (about, e.g., stakeholder requirements) while integrating knowledge of XAI to evaluate its ameliorative prospects (see Chapter 1.4.3).

The papers constituting my PhD project address each of the research questions above. Chapter 2, *Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach* (henceforth CU), addresses RQ1. Chapters 3 and 4, *Explaining AI Through Mechanistic Interpretability* (henceforth XAI-MI) and *Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)* (henceforth RPET), address RQ2. And Chapters 5 and 6, *Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience* (DFB) and *Understanding as a Bottleneck for the Data-driven Approach to Psychiatric Science* (henceforth UB), address RQ3 (see Table 1 for summaries; see Appendix A for clarification about my contributions to the co-authored works).

Chapter 1: Introduction

The remainder of this introduction will consist of the following. In Chapter 1.2, I will provide further exposition of the philosophical and scientific background to the BBP. This will include an explanation of how the BBP arose in light of the history of AI, ML, and DL, and some preliminary remarks on the philosophy of *explanation* and *understanding*. In Chapter 1.3, I will expand on the nature and structure of the BBP. I will provide my own definition of BBP manifestations, detail the constituent concepts of that definition (e.g., stakeholders, desiderata, explainability approaches), and present an original model of how they interact in the BBP. In Chapter 1.4, I will explain how my contributions address research questions 1-3. Finally, I will summarise the contributions of this introduction.

1.2 Scientific and Philosophical Background

The purpose of this section is to provide scientific and philosophical context for what will follow. I begin by introducing the key scientific and technical developments that led the BBP to emerge as an important societal problem. I then turn to a discussion of explanation and understanding, briefly introducing key themes in the philosophical literature before presenting my own approach.

1.2.1 Scientific Background: Artificial Intelligence, Machine Learning, and Deep Learning

The term *Artificial Intelligence* (AI) is highly polysemous (Chollet, 2019; Legg & Hutter, 2007; Wang, 2019). It was coined by the computer scientist John McCarthy in the proposal for a 1956 workshop organised at Dartmouth College. The working hypothesis of researchers attending, as immortalised in a plaque commemorating its 50th anniversary, was that “every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it” (McCarthy et al., 1955). Given the diversity of opinion about what intelligence is, research conducted under the “AI” label reflects a heterogeneous assortment of ideas about how to replicate it mechanically (e.g., Brooks, 1991; Marcus, 2001; Newell & Simon, 1976; Rumelhart, Smolensky, et al., 1986). Since the so-called *deep learning revolution* of 2012, however, AI has become increasingly synonymous with the construction of ANNs trained with machine learning algorithms to approximate target functions (Buckner, 2019; LeCun et

al., 2015). It is this line of research which prompts most worries about the opacity of AI (e.g., Boge, 2022; Burrell, 2016; Zednik, 2021). Here, I present a brief exposition of the relevant technical concepts with a view to clarifying the emergence of the black box problem.

Shortly after the Dartmouth workshop, the term *machine learning* (ML) was popularised by Arthur Samuel, who applied it to develop a checkers-playing system (Samuel, 1959). For Samuel, ML meant programming computers to learn from experience. Unlike traditional computer programs, which implement algorithmic behaviour specified by their programmers (Knuth, 1974), programs developed with ML *learn* their behaviours. The learning procedure is specified algorithmically (via a learning rule) and is constituted by the adjustment of parameters with respect to a performance measure (Chollet, 2021, p. 55). For example, Samuel's checkers-playing system learned to evaluate board positions (in terms of their favourability for one player) by updating the coefficients of a polynomial containing various hand-crafted numerical terms (e.g., ratio of white-to-black pieces).² Samuel's system was the earliest precursor to ML game-playing systems that have achieved superhuman abilities at backgammon, chess, and Go (Silver et al., 2017; Sutton & Barto, 2020). This approach to AI is of enormous historical significance. It enabled, for the first time, the development of systems which learned to solve problems their programmers could not (as presaged by Turing, 1950). The distinct epistemic relation programmers stand in with respect to ML-based systems, as opposed to traditional computer programs, is, in my view, the fundamental source of the BBP.

For the purposes of my dissertation, the most important strand of ML research concerns deep learning and artificial neural networks (Arbib, 1961; Chollet, 2021; Goodfellow et al., 2016; LeCun et al., 2015). Inspired by nerve cells in biological brains, artificial neurons are computational units which produce some degree of activation as a function of the input they receive (this relationship being captured, fittingly, by an *activation function*). For example, McCulloch-Pitts neurons, the first artificial neurons developed, produce an activation of 1 if they receive input above a given threshold and 0 otherwise

² In order for Samuel's system to actually play checkers this learned evaluation function had to be paired with a policy for searching and selecting moves from a partial game tree (e.g., the alpha-beta pruning algorithm).

(McCulloch & Pitts, 1943; note that modern ANNs typically use more expressive activation functions, see Lederer, 2021 for an overview).³ To produce ANNs, artificial neurons are arranged into layered networks consisting of an input layer, an output layer, and, optionally, intermediate layers (also called ‘hidden layers’, see Rumelhart, Hinton, et al., 1986). In the common *feedforward* ANNs, the activation of downstream neurons depends only on the input they receive from upstream neurons.⁴ Such ANNs can thus be represented as directed acyclic graphs (Goodfellow et al., 2016, p. 186). Connections between neurons are equipped with parameters (called *weights*) which serve as multipliers for incoming input. Making these parameters adjustable enables ANNs to (coarsely) simulate biological learning (Zador, 2019).⁵ How an ANN is structured in terms of layers and neural connections is known as its *architecture*. When that architecture involves more than two layers, the network is called a *deep neural network* (DNN), and DL concerns the development and study of such networks. DNNs with non-linear neural activation functions can express complex functions and exhibit sophisticated behaviour.⁶ As DL is a subtype of ML, it involves learning from data via an algorithm (T. M. Mitchell, 2013). One of the distinctive characteristics of DL (as opposed to traditional statistics) is the scale and complexity of the datasets it works with (Hey et al., 2020). Obtaining large enough datasets to learn generalisable patterns (i.e., those that perform well on novel data) is a core challenge of DL (Keogh & Mueen, 2017). As well as selecting the data to learn from, the developers of an ANN typically specify its architecture (including layer-wise connectivity and neural activation functions), its objective function, and its learning procedure.⁷ The objective function (sometimes called a *loss* or *cost function*) specifies the mathematical quantity which the learning procedure aims to maximise or minimise.

³ Mathematically, this behaviour is captured by Oliver Heaviside’s *step function*.

⁴ Other network architectures, such as *recurrent* neural networks, relax this assumption by allowing feedback connections. This enables information to be maintained over time or propagated ‘backwards’ from downstream neurons to upstream ones (Hochreiter & Schmidhuber, 1997).

⁵ Neurons also have *bias* parameters which, when altered, slide their activation function right or left along the x-axis, increasing or decreasing their sensitivity.

⁶ A large technical literature addresses general questions about the approximation properties of particular classes of ANNs. Notably, *universal approximation theorems* show that, in principle, ANNs with hidden layers and non-linear activation functions can “approximate virtually any function of interest to any desired degree of accuracy” (Hornik et al., 1989, p. 359).

⁷ Though, in principle, these elements can also be learned via a higher-order learning procedure (Russell & Norvig, 2020, p. 1203).

For example, in a linear regression problem, the learning procedure may aim to minimise the *mean squared error*, i.e., the sum of squared differences between predicted and true values of the dependent variable.

Central to the DL learning procedure is a learning algorithm, viz., a sequence of computations which adjust the network's parameters with respect to the objective.⁸ The DL learning procedure runs roughly as follows: First, an ANN is initialised with random parameter values (sampled from a statistical distribution such as a Gaussian). These values parameterise a function, f , which the ANN expresses. Then, the ANN operates over a randomly sampled batch of data instances, computing output classifications (for example, assigning class labels to images) or predictions (for example, producing numerical estimates of the value of houses) for each instance. The ANN's performance over this batch is evaluated against the loss function to compute its *error*, expressed as a numerical quantity. Next, the backpropagation algorithm is employed to propagate this error back through the ANN layer-by-layer, computing the partial derivative of the loss function with respect to each parameter. Given this information, each parameter is then updated in the negative direction of the gradient (viz., a vector of all partial derivatives). The magnitude of these updates is defined by a pre-specified *learning rate* (also called *step size*).⁹ This overall process is iterated until the model converges (i.e., when continued training no longer reduces error). The update step of the learning procedure described above is termed *stochastic gradient descent*.¹⁰ For a more detailed explanation of the learning procedure see Chollet (2021, pp. 46–52); for a comprehensive coverage of deep learning, see the textbook by Goodfellow and colleagues (2016).

There are several properties of the DL approach to AI which are important to highlight here. Crucially, the DL methodology continues to reduce the need for domain expertise to build powerful, predictive models (Boon, 2020; Hey et al., 2020). Recall Samuel's

⁸ Technically, the computations involved in training and running ANNs can all be reduced to basic tensor operations (Chollet, 2021, p. 38). These operations can be efficiently parallelised using graphical processing units, a critical fact for the technical feasibility of modern large-scale deep learning.

⁹ In practice, adaptable learning rates with behaviour governed by further algorithms may also be employed to speed up training or increase the probability of convergence to a minimum error value.

¹⁰ It is stochastic because sampling a random batch of data instances renders the gradient computed an approximation of the true gradient. For most real-world applications, the computational cost of computing the true gradient (i.e., using the entire dataset for every weight update) is impractically large.

checkers-playing system which learned to evaluate board positions by weighting hand-crafted features indicative of advantage. Building this system required *some* domain knowledge to select and formalise the features. By comparison, DeepMind’s DL-based Go-playing system, AlphaGo, received “no domain knowledge except the rules” (Silver et al., 2017, p. 354). The features it uses to evaluate board positions are entirely learned from data.¹¹ This example highlights a transition in ML from representation *engineering* approaches to representation *learning* ones like DL (Bengio et al., 2013).¹² Relatedly, the success of ANNs with deep architectures substantiates theoretical arguments that feature hierarchies are critical for complex task performance. As the depth of an ANN is increased, the number of ways it can re-combine information grows exponentially (Eldan & Shamir, 2016). The upshot of this is that deep networks can express functions far more efficiently (i.e., using fewer parameters) than shallow ones (Buckner, 2018, 2019). Finally, it is noteworthy that, from a statistical learning perspective, the success of ANNs has been surprising. In particular, “overparameterised” ANNs, which have more parameters than they observe training instances, have achieved excellent generalisation performance, contradicting the expectations one would derive from classical statistical learning theory (Belkin et al., 2019; Zhang et al., 2021).

Although most of the technical groundwork for DL is decades old (e.g., Ivakhnenko, 1971; see Schmidhuber, 2022 for a comprehensive overview), its perceived potential as an approach to developing AI systems has undergone a dramatic reappraisal in the last decade. Greater availability of computational resources, tailored hardware, and high-quality data led to the development of ANNs exhibiting excellent performance on a range of problems in machine perception (Kell et al., 2018; Krizhevsky et al., 2012), natural language processing (T. Brown et al., 2020; McCann et al., 2017), game-playing (Bakhtin et al., 2022; Perolat et al., 2022; Silver et al., 2017), biochemistry (Jumper et al., 2021), and more besides (Durstewitz et al., 2019; Geer, 2021). This has prompted renewed attention and substantial funding for DL projects, particularly in the technology industry. As Samuel and Turing hoped, ML has been employed to make progress on problems that

¹¹ In the case of AlphaGo, the training data consisted of games of Go played against itself.

¹² Note that the term *representation* is used more liberally in the ML community than in the philosophical community. In ML, intermediate activations of ANN neurons are routinely described as representations, without their having to satisfy the onerous conditions stipulated in the philosophical literature.

stump human experts (Samuel, 1959, p. 223; Turing, 1950, pp. 459–460). However, as was known from the early days of research into ANNs, the parameter values of neural networks interact in complex ways and are not straightforwardly interpretable (Rumelhart et al., 1986, pp. 344–346). Adding to this issue, an important factor in the recent success of DL has been sheer scale (Hoffmann et al., 2024; Kaplan et al., 2020). Language models like GPT3 boast *billions* of trained parameters (T. Brown et al., 2020). And it is the interactions between these parameters which produce the systems’ behaviour. As such, not only are ANNs *prima facie* uninterpretable, they are also extraordinarily complex. Thus, the black box problem for DL rears its ugly head. Just how explainable such systems can be is addressed by RQ2.

1.2.2 Philosophical Background: Explanation and Understanding

The BBP, as I have introduced it, arises due to the unusual epistemic relations which obtain between the developers of DL-based AI systems and the systems themselves. Generally speaking, the programmers of these systems understand them much less well than one would expect through naïve generalisation from the development of other technologies.¹³ Since the ordinary fix for a lack of understanding is to provide an explanation, developing techniques to deliver such explanations is the presumptive goal of the research field of XAI (Buchholz, 2023; Mittelstadt et al., 2019; Rudin, 2019; Zednik, 2021). Thus, clarity of thought around the concepts of *explanation* and *understanding* is of central importance for analysing the BBP. This leads directly to RQ1, which asks how such terms should be understood. Here, I briefly introduce prominent philosophical views on explanation and understanding before presenting my position.

The philosophy of explanation goes back at least as far as Aristotle, whose *four causes* – material, formal, efficient, and final -- are interpretable as complementary explanatory strategies one may pursue in response to a *why*-question (Falcon, 2023; Miller, 2019).¹⁴ More recent theorists of scientific explanation have followed Aristotle’s lead by

¹³ This is not to imply that scientific understanding always precedes technological development. The co-evolution of science and technology is complex and can evolve in multiple ways (Rosenberg, 1982). However, so far as I know, the ability to construct complex technological artefacts without even an *approximate* understanding of how they work is unique to ML-based systems and the BBP.

¹⁴ This is not to make an exegetical claim about what Aristotle intended, but simply to point out the important role his four causes have played in the history of the philosophy of explanation.

considering explanations as answers to *why*-questions (Hempel & Oppenheim, 1948; Van Fraassen, 1980; Woodward, 2003).¹⁵ A core task for philosophers then becomes specifying what kind of structure ought to play this role. Various proposals -- including laws (Hempel, 1965; Nagel, 1982), statistical relationships (Salmon et al., 1971; Van Fraassen, 1980), causal structure (Cartwright, 1994; Woodward, 2003), and mechanisms (Bechtel & Abrahamsen, 2005; Craver, 2006; Salmon, 1984) -- have been defended as either partially or wholly constitutive of explanations.¹⁶ The identification of relevant causes has played a particularly important role in these debates (Craver, 2006; Hitchcock, 1995; Salmon, 1984). In modern parlance, causal explanations can be either *etiological*, specifying the antecedent events that generated an observation, or *constitutive*, specifying how the component parts of a system act to produce its observed behaviour (these are *efficient* and *material* causes, respectively, in Aristotelian terms) (Salmon, 1984). Other theorists have additionally argued that non-causal explanations, such as mathematical explanations, also play an indispensable role in science (Andersen, 2018; Lange, 2017). Finally, some theories of explanation move away from the domain of science. For example, action theorists are concerned with explanations for (intentional) actions (Setiya, 2003). Here, explanations often take the form of motivating reasons, i.e., (rational) considerations in light of which an agent acts (Glock & Schmidt, 2021; Hieronymi, 2011).¹⁷ The conditions AI systems would need to meet for such explanations to apply to their behaviour is unclear (Baum et al., 2022; Swanepoel, 2021).

Which theory of explanation is best for the BBP? In line with the heterogeneity of BBP manifestations, I believe a pluralistic approach is required. AI systems can be usefully conceptualised as mathematical models, objects of scientific interest, technical artefacts, and agents that take actions in the world (Eden, 2007; Newell & Simon, 1976). Thus, depending on the circumstances, various explanatory strategies may be required to render their behaviour intelligible, often in combination (S. D. Mitchell, 2002). In line

¹⁵ See Faye (2014a) for a dissenting view.

¹⁶ Note that these proposals are not mutually exclusive; many popular views of explanation involve combinations of the listed ingredients.

¹⁷ Motivating reasons are typically distinguished from normative reasons, which are reasons objectively counting in favour of a given action (but which need not be considered by the agent). Note that motivating reasons may not be identical to explanatory reasons (Alvarez, 2009).

with this, Erasmus and colleagues argue persuasively that multiple models of explanation from the philosophy of science can all be applied, in principle, to ANNs (Erasmus et al., 2021). Note too that while my discussion has primarily focused on scientific explanation, introducing details which may not be relevant to all of the explanatory ambitions of XAI (e.g., developing explanations for laypeople), some philosophers have argued that everyday explanations are continuous with scientific ones (Wilkenfeld, 2014; Woodward, 2003). On this view, explanations of ordinary events are not fundamentally different to scientific explanations.¹⁸ That said, it is clear that practical choices about detail, format, and relevance diverge for scientific and everyday explanations (Lombrozo, 2016; Miller, 2019; Vasilyeva et al., 2017).

A final distinction in the philosophy of explanation which is important to mention here concerns the relata between which explanatory relations are said to hold. Traditional philosophical approaches typically treat scientific explanations as a two-part relation between explanans and explanandum (e.g., Hempel, 1965).¹⁹ Pragmatic theories, in contrast, introduce the *recipients* of explanations as ineliminable components, arguing that explaining is a fundamentally contextual, communicative endeavour (Faye, 2014b; Nyrup & Robinson, 2022; Potochnik, 2016).²⁰ This distinction affects how explanations should be evaluated, since the value of an explanation for recipients typically depends on at least some features other than its relation to the explanatory target (e.g., intelligibility). In line with the pragmatic view, some philosophers of science have emphasised that explanations can only be judged relative to particular epistemic perspectives governed by local epistemic norms (Kästner, 2018; Kästner & Haueis, 2021; Wimsatt, 1994). Along similar lines, theoretical views of explanation in the cognitive and developmental sciences foreground the requirements and capacities of recipients (Gopnik, 2000; Keil, 2006; Lombrozo, 2006). Common to these views is the idea that explanations serve practical functions, and ought to be assessed, at least partially, on

¹⁸ Obviously, what exactly this essential similarity consists in depends upon one's preferred theory of explanation.

¹⁹ A further possibility, the *ontic* view, argues that the relata of explanations are different parts of the structure of reality, such as causes and their effects (Craver, 2007, p. 27).

²⁰ Nyrup and Robinson go one step further in their *explanatory pragmatism* framework by explicitly including an *explainer*.

that basis. Given that the BBP is a practical problem, pragmatic approaches to explanation seem particularly appropriate for the subject of XAI.²¹ However, as I will explain below, I prefer an approach which treats explanation as derivative of *understanding*.

Understanding is a topic that received less philosophical attention than explanation in the 20th century (de Regt, 2019; Wilkenfeld, 2014). Since then, however, the philosophy of understanding has blossomed in both epistemology and the philosophy of science (Baumberger et al., 2016; Grimm, 2019). One key idea arising from this literature is the differentiation of understanding into *explanatory* and *objectual* varieties (Elgin, 2007; Kvanvig, 2003). Explanatory understanding is understanding *why* something is the case and is often accompanied with a particular contrast (e.g., why p and not q). Objectual understanding is understanding a domain of inquiry in terms of how the elements comprising it relate to one another, often in terms of causal and/or constitutive structure. The distinction between explanatory and objectual understanding can be leveraged to ground a distinction between understanding and knowledge. Arguably, explanatory understanding is propositional, and thus just a special case of knowledge (viz., knowing why p and not q) (pace Hills, 2016). However, objectual understanding goes beyond knowledge and involves cognitive skills, such as the ability to combine information and draw inferences (Baumberger et al., 2016; Grimm, 2019; Hills, 2016).²² This brings out another important theme of philosophical work on understanding: its tight connection to practical abilities (CU and DFB both build on this idea). Intuitively, we ascribe understanding to others just in case they are able to perform feats characteristic of one who understands (Wilkenfeld, 2013; Wilkenfeld & Hellmann, 2014). Various such feats have been proposed, including the ability to make predictions, construct explanations, and control phenomena of interest (de Regt, 2019; de Regt & Dieks, 2005). Note that while

²¹ As a pluralist, I maintain that multiple philosophical accounts of explanation may be discursively useful. For example, Hempel and Oppenheim's deductive-nomological model of scientific explanation, while not pragmatic, elucidates an important aspect of how scientific theories explain.

²² A proponent of the view that all understanding is knowledge can, of course, appeal to the idea that these cognitive skills are just different types of knowledge (Khalifa, 2017). In my view, the key point is that objectual understanding depends upon cognitive skills that go well beyond storage and retrieval of semantically encoded information. Whether or not those skills are classified as knowledge is then a semantic question.

cognitive skills likely play a crucial role in such feats, practical abilities (e.g., controlling a phenomenon) are not always reducible to cognitive skills, as they plausibly require non-cognitive support in addition.²³ The association of understanding with practical abilities immunises it from worries that understanding is too subjective a phenomenon to warrant philosophical consideration (Hempel & Oppenheim, 1948). This worry is natural if understanding is construed purely as an “a-ha” feeling (since such a feeling may accompany misunderstanding) but unwarranted if understanding is given a more substantive interpretation (de Regt, 2009; P. Lipton, 2008; Wilkenfeld, 2014).

Philosophers interested in explanation and understanding often prioritise one over the other. The reason for this is that, on many views about the nature of the relation between the two concepts, an account of one suffices to capture the relevant aspects of the other (Khalifa, 2017). For example, *understanding* is often construed as a by-product of receiving or grasping an explanation (Hempel, 1965; Strevens, 2013). On this view, an account of what explanations are, along with an analysis of what it means to grasp one, suffices as an analysis of understanding. The complementary approach, which I pursue in this dissertation, begins with an account of understanding (e.g., Wilkenfeld, 2013; Van Camp, 2014). One can then define an explanation as anything that, under the right conditions, produces understanding (Wilkenfeld, 2014). In my view, the fact that understanding is typically downstream of explanatory information and closer to the desiderata we ultimately care about in the BBP (e.g., safety, trust etc...) speaks in favour of an understanding-first strategy. This renders explanation a derivative notion (for related arguments see Nyrup & Robinson, 2022; Páez, 2019). Notably, given the diverse forms of information that might induce understanding (see Langer et al., 2021, pp. 9–11 for examples), understanding-first approaches generally render explanations a broader and more heterogeneous category than sometimes supposed in the philosophical literature. To retain the distinction between this broader perspective and more restrictive views of explanation, I adopt the term *explanatory information* (taken from Langer et al., 2021) to refer to whatever information might induce understanding.

²³ Candidates for such support include relevant knowledge, experience, physical abilities (e.g., stamina, motor control), and more. Theorists may differ as to which of these candidates ought to be considered cognitive. Though I am sympathetic to an inclusive view, discussing this matter further here is out of scope.

Understanding is a unifying theme of this dissertation, and much more will be said about it in the articles following this introduction (particularly CU, DFB, and UB). However, at this point, I will briefly clarify how I think about understanding. First, as with explanation, I endorse a pluralist approach to understanding. This means that I believe multiple, distinct construals of what understanding is can play valuable roles in particular scientific, practical, and philosophical contexts. For example, one might construe understanding as a special kind of cognitive state (e.g., maximally comprehensive knowledge, Kelp, 2015); or, one might say that understanding *constitutively* requires whatever else allow the cognitive state to manifest in observable practical abilities under a range of ordinary (human) conditions (e.g., Hills, 2016; Wilkenfeld, 2017).²⁴ Plausibly, both of these notions (as well as many others) can play useful roles in certain discursive contexts and no knock-down argument speaks for one over the other as a general analysis of understanding. Which construal of understanding is preferred in a given context will depend on details about one's discursive aims (see Carruthers, 1987 for discussion of this form of conceptual pragmatism). Since this dissertation is concerned with the BBP, practical consequences (i.e., the satisfaction of desiderata) are of particular importance. I take this to support the inclusion of practical abilities in the construal of understanding most fitting for this context. To clarify, then, I make no claims about *how exactly* such abilities are brought about by underlying constructs like cognitive skills, knowledge, or motivation. Indeed, it is a feature of endorsing an abilities-based view that one can stay neutral on such questions. This line of argument is expanded in detail in CU. There are further noteworthy features of the view of understanding I adopt in this dissertation. For one, I take it that understanding is *graded*. That is, one can understand something to a greater or lesser degree (Baumberger, 2019). As a corollary, binary attributions of understanding (e.g., "Agatha understands chess") involve context-relative thresholds and do not have general, determinate truth-conditions (Wilkenfeld, 2017). The graded nature of understanding is discussed further in DFB. I also take a broad view of

²⁴ These two positions echo longstanding debates in the philosophy of linguistics and cognitive science about the distinction between competence and performance (see, e.g., Nelson, 1978; Trützsch-Wijnen, 2020; Firestone, 2020). Narrower construals of understanding limit it to an idealised set of cognitive structures (hypothesised to underlie competence), while broader construals include whatever is necessary to enable performance.

what the *object* of understanding can be. Understanding is a two-place predicate. That is, *someone* understands *something*. Philosophers have observed that linguistically acceptable objects of this locution are varied (Hills, 2016; Wilkenfeld, 2014) and I am inclined to think most linguistically acceptable objects are philosophically acceptable. Generally speaking, so long as a person's ability to bring about desired consequences through interaction with something is sensitive to the person's cognitive skills and associated practical abilities, treating that thing as the object of understanding is reasonable.²⁵ For the purposes of this dissertation, the relevant object of understanding will typically be some aspect of an AI system, though in UB the object of *patient understanding* is a patient's own mental states.

Further questions often addressed by full-blown accounts of understanding concern its *factivity*, viz., whether the content of understanding needs to be true, and its relation to *belief* and *knowledge* (see, e.g., Baumberger, 2019; Doyle et al., 2019; Elgin, 2017; Grimm, 2006; Kvanvig, 2003). I cannot discuss these matters in detail here, but I will briefly comment on how the abilities-based view relates to them. Crucially, if understanding is associated with abilities and *whatever* underlies them, then it often involves non-propositional content which is not truth-apt (Baumberger, 2019). This immediately implies that understanding cannot be factive in any straightforward sense (i.e., identical to a grasp of true propositions). Despite this, it is clear that "understanding somehow answers to facts" (Elgin, 2007, p. 33). In other words, the requirement that cognitive structures suffice for relevant abilities (e.g., inference, control) functions as a strong constraint on the content of understanding. Whether this is a non-factivist or quasi-factivist position is arguably a matter of taste (Gurova, 2022). Another consequence of understanding involving non-propositional content is that it cannot be reduced to knowledge or a set of beliefs (assuming these must be propositional). However, it is plausible that understanding of *some* objects (especially propositions) requires particular knowledge or beliefs of a certain sort. Finally, I emphasise again that I only take the abilities-based view to constitute the best construal of understanding for the context of the BBP. Thus, these implications are also restricted to that domain.

²⁵ Note that this is a sufficient condition, not a necessary one. As such, understanding of an object one cannot possibly control (e.g., a past event) is not ruled out.

1.3. The Black-Box Problem

With the scientific and philosophical background outlined, let me now turn to the black box problem itself. The BBP is a complex beast comprising numerous moving parts. I begin this section by presenting my own simple definition. I then enumerate and describe the critical conceptual terms, including an original taxonomy of explainability approaches. Finally, I introduce a novel model of the BBP which serves to clarify its structure and situate my research contributions.

1.3.1 Key Concepts

BBP manifestations: A definition

For the BBP to manifest, the opacity of an AI system must prevent the satisfaction of some desideratum (Zednik, 2021). For example, consider the following case. An opaque AI system is entrusted with granting bank loans to applicants. Alice the applicant wants the decision procedure to be fair. However, the opacity of the system prevents Alice from assessing whether her application was rejected fairly or not. To solve this particular manifestation of the BBP we require an answer to the question *why did the AI system refuse to grant Alice a bank loan?*²⁶ As above, I will use the term *manifestation* to refer to instances of the BBP. I begin by providing a definition:

BBP Manifestation: A set of human-system interactions in which the opacity of an AI system prevents the satisfaction of a stakeholder-held desideratum.

To a first approximation, the variables required to specify a BBP manifestation are the affected parties (stakeholders), their values and goals (desiderata), the AI system involved and its opacity/explainability properties, and the relevant human-system interactions. As with Alice, above, for the BBP to manifest at least one stakeholder desideratum must not be satisfied *due to* the opacity of an AI system. This means the

²⁶ Of course, discovering that the AI system discriminated on the basis of a protected attribute would not satisfy the desideratum for fairness. However, assuming accurate explanatory information was provided, the BBP would no longer obtain, because opacity would cease to play the relevant preventative role. Another way to put this is that Alice's *epistemic* desideratum (viz., to assess system fairness) was satisfied, but her *substantial* desideratum (viz., to be treated fairly) was not (see Langer et al., 2021, pp. 7–8 for further discussion).

faulty performance of a perfectly understood system is not a BBP manifestation. The aim of XAI, then, is to address problematic sources of opacity by producing explanatory information that increases a stakeholder's understanding of the system (though not necessarily the same stakeholder whose desideratum is prevented by opacity). Increased understanding affects the course of human-system interactions and thus potentially facilitates desiderata satisfaction.²⁷ In Chapter 1.3.2 I present my conceptual model of the BBP. Before that, I say more about each component of the model.

Stakeholders

BBP contexts involve numerous kinds of stakeholders (Buchholz, 2023; Langer et al., 2021; Tomsett et al., 2018). These include developers, deployers, operators, users, data subjects, decision subjects, regulators, and, plausibly, society itself.²⁸ Stakeholder types are individuated by the role(s) they play in BBP manifestations and the kinds of human-system interactions they enter into. Developers are people who produce an AI system. Deployers are people who choose to put an AI system into use. Operators are people who control the use of an AI system. Users are people who interact with a deployed AI system. Data subjects are people whose data is processed by an AI system. Decision subjects are people about whom an AI system makes a decision. Regulators are people who ensure the operation of an AI system complies with norms, whether legal, ethical, or otherwise. Finally, how AI is used (and that its benefits are widely distributed) is a subject of general interest in a democratic society (Seeger et al., 2023). Stakeholder roles are not mutually exclusive; individuals often occupy multiple roles (e.g., developer, deployer, and operator) simultaneously. The roles listed above are probably not exhaustive, but they cover those most commonly discussed in the literature.

Stakeholders differ in numerous important ways. Since the BBP concerns epistemic barriers to desiderata satisfaction, *background knowledge* is particularly relevant. The background knowledge and skills stakeholders possess strongly constrains the kinds of explainability approaches that will plausibly be useful to them (Belle & Papantonis, 2021;

²⁷ Note that my definition captures cases in which reducing opacity is *necessary* for desiderata satisfaction. It does not imply that reducing opacity *suffices* for desiderata satisfaction.

²⁸ One could also consider further composite entities such as nations, companies, governments, and demographic groups. Though valuable, analysis at this level of description is beyond the scope of this dissertation.

Fleisher, 2022; Langer et al., 2021; Páez, 2019). For example, developers of AI systems will typically possess more technical knowledge than users or data subjects (Guidotti et al., 2021). Thus, developers may benefit from explanatory information of greater complexity and volume than stakeholder groups with less domain knowledge. Relatedly, stakeholders likely differ in terms of their *motivation*, including a propensity to engage with explanatory information (even for a fixed level of background knowledge) and the time they are willing to invest in improving their grasp of a system (Belle & Papantonis, 2021). Further, stakeholders may have access to different quantities and kinds of *resources* to aid them in ameliorating problems arising from the opacity of AI systems. Relevant resources could be financial, human, or technical (e.g., specialised hardware). I subsume stakeholders' background knowledge, motivation, and resources under the term *stakeholder profile*. Stakeholder profiles can be ascribed to roles or individuals depending on the generality of an analysis.²⁹

Desiderata

Desiderata are the things stakeholders want from the use of AI (broadly construed). In accordance with the diversity of stakeholders, desiderata are highly heterogeneous. Examples include safety, performance, responsibility, transparency, trustworthiness, and fairness (Amodei et al., 2016; Barredo Arrieta et al., 2020; Guidotti et al., 2018; Kästner et al., 2021; Langer et al., 2021; Schoeffer et al., 2024). Desiderata vary considerably in terms of their specificity: from the satisfaction of particular technical specifications to the broad promotion of vaguely defined values (Köhl et al., 2019; Munn, 2023). In some cases, desiderata may be strict requirements, as in the case of compliance with legal statutes. The European AI Act constitutes the most extensive attempt to date to codify desiderata in legal terms (Artificial Intelligence Act, 2024; see Novelli et al., 2023 for discussion). In other cases, however, desiderata go beyond legal necessity. They may include accordance with ethical principles, such as justice or fairness, or with particular schools of moral philosophy, such as utilitarianism or virtue ethics (Boddington, 2023; Serafimova, 2020). However, desiderata may also encompass (plausibly) morally neutral preferences, such as a desire to understand for its own sake

²⁹ Naturally, stakeholder groups are not monoliths and generalisations captured by stakeholder profiles may fail to hold in particular cases.

or adherence to particular aesthetic principles (Chiarella et al., 2022; Hullman et al., 2023). Distinct desiderata may be possessed by individual stakeholders. For example, a doctor employing an AI decision-support system may value ease of use, while the doctor's patient may value privacy. However, desiderata can also be shared, as both the doctor and patient may value the trustworthiness of the system. Desiderata may also be held by composite entities. For instance, a government watchdog may demand a particular level of transparency from a company deploying an AI system (Floridi et al., 2018). Often, however, such desiderata can be ascribed to individuals when considering concrete cases, e.g., the employee(s) responsible for ensuring the company's compliance with the watchdog's desideratum. Note that some desiderata have little to do with opacity (Ananny & Crawford, 2018; Durán & Jongsma, 2021; Krishnan, 2020). Thus, recognising when it is *opacity*, in particular, that impedes desiderata satisfaction is an important step in identifying BBP manifestations.

AI systems

The BBP manifests when the opacity of an AI system prevents the satisfaction of a desideratum. But what exactly constitutes an *AI system*? As discussed in Chapter 1.2.1, there is no scientific or philosophical consensus on which range of systems count as AI (Bringsjord & Govindarajulu, 2024). As such, I am inclined to follow usage and accept whichever systems are referred to as “AI” by relevant epistemic communities as, de facto, AI systems. This policy results in liberal criteria (see, e.g., Russell & Norvig, 2020, Section 2.1).³⁰ However, it is important to emphasise that the intensity of research around the BBP has been motivated specifically by the success and prevalence of ANNs and other ML-based systems (Rudin, 2019; Zednik, 2021). As discussed in Chapter 1.2.1, what distinguishes these systems from traditional computer programs and symbolic approaches to AI is that their internal structure (viz., functional organisation encoded in parameter values) is developed through an automated learning procedure. Approaches to AI which exploit such processes sever the dependency between human knowledge of how a problem might be solved and the resulting implementation of system behaviour.

³⁰ Foundational figures like McCarthy and influential practitioners like Russell and Norvig both champion liberal and inclusive construals of *AI system/agent* as anything which can be modelled as producing outputs in selective response to inputs (McCarthy, 1979; Russell & Norvig, 2020).

Thus, while the BBP can and does manifest for traditional computer systems (see Mann et al., 2023 for examples), most philosophically novel aspects of the BBP concern how it manifests for ANNs and other DL-based systems (Boge, 2022).³¹ In accordance with these considerations, I focus on the explainability of DL-based systems in XAI-MI, DFB, and UB. However, other systems may also feature in BBP manifestations.³²

Explainability approaches

Thus far I have said little about XAI itself. The aim of XAI is to render AI systems explainable. I take this to be equivalent to enabling understanding by reducing opacity (about which more later).³³ There are multiple ways to achieve that aim and assessing their potential is the subject of RQ2. One crucial distinction is between XAI *methods* and XAI *approaches*. XAI methods are computational procedures which can be applied to AI systems to produce explanatory information. XAI approaches constitute a more general class, including alternative means of reducing opacity (see Table 2 for a summary). The *method-centric* approach has dominated XAI so far and I will discuss it further below. In addition to that approach, three others are critical to identify.³⁴ First, the project of *mechanistic interpretability* (MI), which is the subject of XAI-MI, combines the systematic application of XAI methods with the methodology of scientific discovery to try and uncover the functional organisation of trained AI systems (e.g., Elhage et al., 2021; Olah et al., 2018). Second, the approach of *ante-hoc explainability* (also referred to as transparent model design, white box models, or inherent interpretability) involves ensuring that a system is explainable by making particular choices during the design and

³¹ The challenge of the BBP for traditional computer systems is essentially the same as the epistemic challenges raised by other complex technological artefacts.

³² Note that many AI systems are hybrid and modular, involving the integration of ANN-components, symbolic reasoning modules, and traditional control structures (Garcez & Lamb, 2023; Silver et al., 2017; Wermter & Sun, 2000).

³³ There are many possible distinctions to be made between near-cognates like explainability, interpretability, transparency, and understandability (for possibilities, see Erasmus et al., 2021; Fleisher, 2022; Z. C. Lipton, 2018). I avoid detailing such distinctions here to avoid an explosion in conceptual complexity.

³⁴ I am not aware of any other prominent approaches to reducing the opacity of AI systems. However, given the enormous volume of academic work on AI systems, I cannot be confident that these approaches exhaustively partition the space of possibilities.

development process (Molnar, 2022, Chapter 5; Rudin, 2019).³⁵ Ante-hoc explainability contrasts with *post-hoc* explainability, a term referring to opacity reducing procedures that occur *after* an AI system's development (this classification would subsume MI and much of the methods-centric approach). The potential of the ante-hoc approach is discussed in RPET. Third, it is also worth highlighting the *theoretical* approach to understanding AI systems.³⁶ This approach treats deep learning systems as mathematical objects and attempts to understand various aspects of their development and behaviour on that basis (e.g., Bronstein et al., 2021; Prince, 2023; Roberts & Yaida, 2022). As a caveat, though these XAI approaches are distinguishable, they have fuzzy boundaries and interact with one another.³⁷ In my view, some understanding of the role of each approach is necessary for having an accurate view of the XAI landscape.

The majority of research in XAI has concerned the development of *methods* (Belle & Papantonis, 2021; Guidotti et al., 2018; Marcinkevičs & Vogt, 2023; Rätz, 2023). Methods are computational procedures that can be applied to AI systems to produce explanatory information (Buchholz, 2023 calls these 'instruments'). To a first approximation, the method-centric approach addresses the BBP by developing methods that deliver explanatory information to stakeholders *in situ* (i.e., during the course of direct human-system interactions) (Belle & Papantonis, 2021; Buchholz, 2023; Fleisher, 2022). While I believe the complementary approaches described above are also important for analysing the BBP (a theme I discuss further in XAI-MI), it is true that methods are central to XAI.³⁸ There are many different XAI methods and many ways to categorise them (I can

³⁵ One might reason that the ante-hoc explainability approach is not an XAI approach at all. After all, the approach involves deliberate avoidance of the opacity that characterises BBP contexts. To be sure, in some cases one might replace an extant opaque system with an inherently interpretable one. However, fans of ante-hoc explainability generally favour avoiding deployment of opaque systems in the first place. To capture such cases, one could alter my definition of BBP manifestations to be defined by a *risk* of opacity preventing desiderata satisfaction. Practically speaking, not much turns on this issue.

³⁶ Theoretical research on ML, DL and AI is not usually considered part of XAI. However, a comprehensive view of how greater understanding of AI systems might increase satisfaction of desiderata must include the theoretical approach. I believe the relationship between the theoretical approach and XAI is understudied, but this is not a subject that I address in this dissertation.

³⁷ For example, theoretical understanding is useful for developing effective methods, effective methods are crucial for mechanistic interpretability, and high-quality mechanistic interpretability enables the design of ante-hoc explainable systems.

³⁸ Indeed, several authors (e.g., Buchholz, 2023; Fleisher, 2022; Zhou et al., 2021) construe the goal of XAI narrowly as the development of methods. Though I opt to call all the explainability approaches in Table 2

TABLE 1: OVERVIEW OF EXPLAINABILITY APPROACHES

Explainability Approach	Description	Example	Relevant Publications
Method-centric XAI	Developing computational methods to deliver explanations to stakeholders <i>in situ</i>	Counterfactual explanations provide actionable information to subjects of AI bank loan decisions	(Karimi et al., 2022; Ribeiro et al., 2018; Wachter et al., 2017; Zhou et al., 2021)
Mechanistic Interpretability	Applying coordinated discovery strategies to reverse engineer the functional organisation of trained ANNs	Feature visualisation reveals multimodal neurons in visual classification models, enabling hand-designed adversarial examples	(Elhage et al., 2021; Geiger et al., 2022; Goh et al., 2021; Olah et al., 2018)
Ante-hoc explainability	Developing AI systems which are inherently interpretable to human stakeholders	Using inherently interpretable AI architecture (e.g., decision trees) ensures rationale for medical decisions is understood by patient	(Dierckx et al., 2023; Ghassemi et al., 2021; Rudin, 2019; Sarkar et al., 2022)
Theoretical explainability	Theorising about deep learning and AI systems as mathematical objects	Understanding role of overparameterisation in DL enables training of far smaller models with equivalent performance	(Bronstein et al., 2021; Frankle & Carbin, 2019; Prince, 2023; Zhang et al., 2021)

only provide a coarse overview, for details see Guidotti et al., 2018; Marcinkevičs & Vogt, 2023; Speith, 2022). *Local* methods produce information pertaining to specific inferences made by an AI system. For example, a *saliency map* might highlight pixels relevant to an image’s classification as dog (e.g., Ribeiro et al., 2016; Kindermans et al., 2019). *Global* methods produce information pertaining to the system’s behaviour as a whole. For example, a *surrogate model* approach might involve transforming an ANN into

“XAI approaches” here, the critical point is the existence of alternative approaches for reducing opacity and solving the BBP. Whether or not those alternatives are labelled part of XAI is largely a semantic issue.

an interpretable form such as a decision tree (e.g., Bastani et al., 2019). *Model-agnostic* methods are those which can be applied to any model (e.g., Ribeiro et al., 2018). For example, the LIME method produces a simplified surrogate model that approximates the local behaviour of any system (Ribeiro et al., 2016). *Model-specific* methods can only be applied to particular types of systems. For example, *feature visualisation* is an XAI method that generates an image maximising the activation of particular ANN components (e.g., neurons) (Nguyen et al., 2019; Olah et al., 2017). Methods can be further demarcated by their means of functioning (e.g., perturbation-based, gradient-based), the result they produce (e.g., data instances, measures of feature relevance), and the format of their output (e.g., text, image) (see Speith, 2022 for discussion). I will argue in what follows (especially Chapters 3 and 4) that XAI methods and other XAI approaches have particular strengths and weaknesses and are suited to different ends.

Opacity

XAI aims to bring about the satisfaction of desiderata via the reduction of *opacity*. But what exactly is opacity? One useful definition takes opacity to be any property of a potential object of understanding (broadly construed, as per Chapter 1.2.2) that functions as a *barrier to understanding* (Beisbart, 2021; Mann et al., 2023). On this view, opacity is some state of a target object that, were it different in relevant ways, would enable understanding. It should be clear that viewing opacity this way makes it agent-relative (Humphreys, 2009). If understanding is a state predicated of individuals with respect to objects, then what constitutes a barrier to understanding will depend on properties of the individuals in question (to put it in the terminology introduced above, it depends upon the relevant *stakeholder profiles*). Generalisations about the opacity of particular systems, e.g., “ANNs are opaque”, are thus relational statements with implicit indirect objects (viz., those to whom the ANNs are opaque). Consequently, such generalisations are apt to change in line with our ability to reduce the barriers which make the statements (approximately) true.³⁹ This is discussed further in DFB. Recall that I construe understanding as, roughly, a person’s practical abilities (e.g., predicting,

³⁹ Humphreys (2009) invokes the notion of *essential* opacity to capture features of objects (*epistemically relevant elements*, in his terminology) that will *always* elude human understanding. The question of whether certain aspects of sufficiently complex AI systems are essentially opaque is interesting but I cannot address it here.

explaining, intervening) with respect to an object. In the context of XAI, the relevant object is typically an AI system (or some part or behaviour of such a system) but may also include the broader sociotechnical system in which an AI system is embedded (see, e.g., Vredenburg, 2022). Different XAI approaches can be understood as targeting different barriers to understanding. For instance, one barrier is the absence of actionable information for subjects of AI decision systems. This is addressed by work on algorithmic recourse (e.g., Karimi et al., 2022). Another barrier is the absence of detailed explanatory information for the behaviour of extant systems. This is addressed by ongoing projects of mechanistic interpretability (e.g., Bricken et al., 2023). Yet another barrier is the complexity of deep learning systems like ANNs. This is addressed by the development of ante-hoc explainable systems (e.g., Felzmann et al., 2020). As with XAI methods, theorists have worked to categorise and catalogue the opacity that arises in BBP manifestations and beyond (see Boge, 2022; Burrell, 2016; Facchini & Termine, 2022; Mann et al., 2023).

Human-system interaction

Manifestations of the BBP involve human-system interactions. A minimal human-system interaction involves some activity of an AI system, some activity of a human stakeholder, and some form of dependence between the two (Janlert & Stolterman, 2017; Schleidgen et al., 2023).⁴⁰ Interactions can be characterised by three key parameters. First, they may be *direct*, such as using Google Maps to choose one's route home, or *indirect*, such as being guided through medical treatment options by a doctor whose opinion is informed by an AI support tool.⁴¹ Notably, given the prevalence of AI and computer systems in modern society, indirect interactions are ubiquitous and may be obscured (i.e., the involvement of a computer system is not apparent). Interactions can also be *continuous*, such as operating a level 3 autonomous vehicle, or *discrete*, such as generating a single image with a generative AI tool (Faconti & Massink, 2000). Finally, interactions also differ

⁴⁰ Note that I have tailored this definition of human-system interaction for my analysis of the BBP. It is broader than other definitions proposed in the academic discipline of *human-computer interaction* (see, e.g., Janlert & Stolterman, 2017). Narrower definitions may be suitable for particular epistemic goals, e.g., investigation of design principles ensuring user satisfaction.

⁴¹ Note that these definitions do not track how *direct* is used in HCI (see, e.g., Shneiderman & Maes, 1997; Winograd, 2006)

with respect to the degree of *control* the human exerts over the system (and vice versa). The case in which the system (human) is fully autonomous constitutes one end of the spectrum, the case in which the system (human) is fully controlled constitutes the other (Shneiderman & Maes, 1997). BBP manifestations typically concern intermediate cases.

Manifestations revisited

Recall that I defined a BBP manifestation as a set of human-system interactions in which the opacity of an AI system prevents the satisfaction of a stakeholder-held desideratum. Having explicated the notions the definition employs, I wish to clarify several issues. First, real-world BBP manifestations can be massively complex, involving the interdependent action of numerous stakeholders, computer systems, and further intermediaries. Further, opacity can cause problems at numerous points along the lifecycle of an AI system, including development, testing, and deployment (De Silva & Alahakoon, 2022; Deck et al., 2024). Thus, the point at which a particular desideratum fails to be satisfied may be far removed, temporally, from the point at which opacity played a key role. For example, consider a decision subject automatically rejected from a job application by an opaque AI system in virtue of an irrelevant feature (say, the name on their CV) correlating with poor performance amongst past employees. Here, opacity during the development and evaluation of the system prevents desiderata satisfaction during its deployment. Second, as the foregoing example shows, the stakeholder whose desideratum is at stake in a given BBP manifestation may differ from the stakeholder whose understanding could ameliorate the situation. In the job application case, the situation may have been avoided had the developers or deployers of the AI system understood its decision-making procedure, but it was the decision subject whose desideratum (fair treatment) was not satisfied. Third, BBP manifestations are common. That *somebody* understanding *some aspect* of an AI system better could improve a desiderata-related outcome is a weak condition. Indeed, I conjecture that this condition holds for most practical applications of AI systems. However, BBP manifestations vary considerably with respect to their importance. Accordingly, analysis should (and does) focus on manifestations of particular moral, societal, and scientific significance. Finally, BBP manifestations also vary with respect to their identifiability. In some contexts, the dangers of opacity are striking and have been extensively discussed (e.g., the operation of autonomous

vehicles; the use of decision support systems in medical settings; the use of deep learning in science). However, there are many contexts in which AI systems play non-obvious roles (from the perspective of some stakeholders).⁴² Thus, though my definition of BBP manifestations is broad, identifying particular cases is an important challenge.

1.3.2 A Model of the BBP

With the parts that compose BBP manifestations explicated, I now introduce my own conceptual model of the structure of the problem (Fig. 2).

FIGURE 2: A MODEL OF THE BLACK BOX PROBLEM

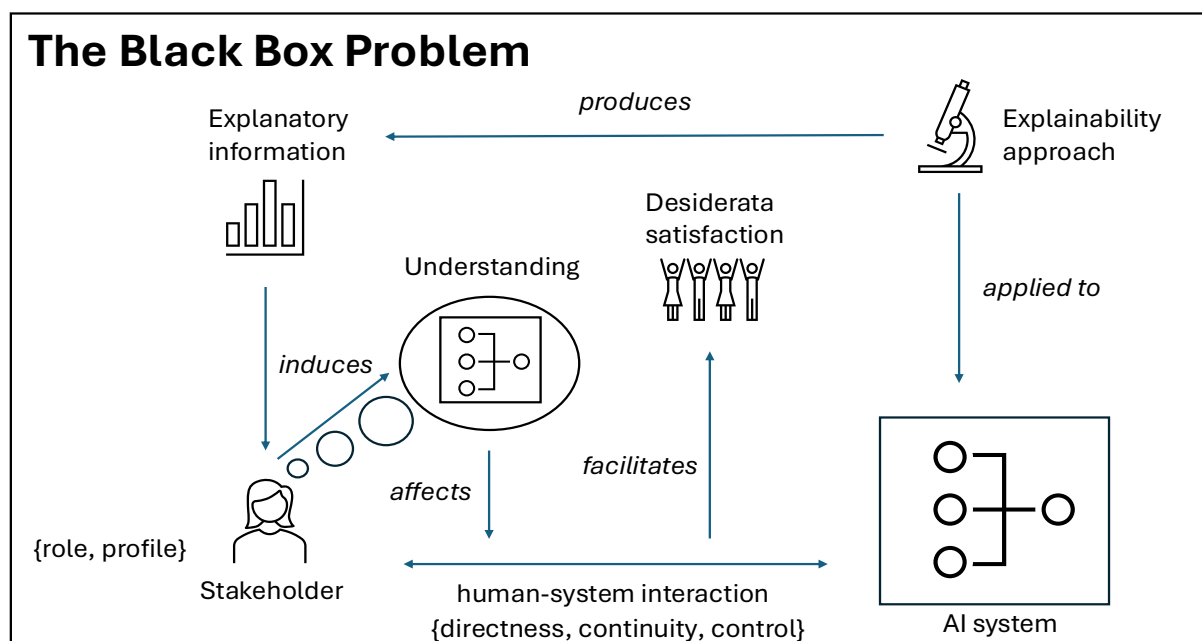


Figure 2 depicts a simple model of the BBP. The bottom left of the diagram depicts a stakeholder who interacts with an AI system in the bottom right. The stakeholder understands the AI system and/or its behaviour to a certain degree, represented by the mental model in the thought bubble. This degree of understanding affects the stakeholder's interaction with the AI system, potentially facilitating the satisfaction of stakeholder-held desiderata. Explainability approaches (top right) can be applied to the AI system to produce explanatory information. This information induces understanding in the stakeholder, affecting their ability to satisfy desiderata. The properties in braces represent details that must be specified about stakeholders and their interactions with AI systems to adequately characterise BBP manifestations (see Chapter 1.3.1).

⁴² This may be due to the indirectness of the interaction, deliberate misrepresentation by an interested party, technical illiteracy, or a combination of factors.

For visual clarity, this model is simplistic. It does not capture everything that might be relevant for a BBP manifestation. Notably, the model depicts a single stakeholder entering into relations with all the other elements. Often, multiple stakeholders will enter into different relations with the other elements of a given manifestation. For example, consider a medical support system which a doctor uses when evaluating treatment options for a patient. Here, both the doctor (user/operator) and the patient (decision subject) are stakeholders. However, depending on the nature of the inquiry, one may also consider developers, deployers, and regulators. Indeed, these stakeholders may be the most effective targets for ameliorative interventions. For example, developers may be responsible for designing the system such that it presents information in a way that accurately reflects uncertainty about the costs and benefits of particular treatment courses.⁴³ To be clear, the effects of such development decisions on desiderata satisfaction may route through the understanding of the doctor and the patient. However, a useful model of the BBP should encompass all plausible targets for intervention. Thus, stakeholders engaging with upstream barriers to understanding must be considered part of BBP manifestations.

Let us now return to the three organising questions of my dissertation and recontextualise them with the model above in mind (see Chapter 1.4 for graphical guides). RQ1 concerns how *understanding* should be construed in the context of the BBP. The model highlights the significance of this question. The outcomes we ultimately care about are represented by *desiderata satisfaction*. In the BBP model, by construction, *understanding* is the epistemic construct which affects, via human-system interactions, the degree to which desiderata are satisfied. Thus, understanding is a crucial target for philosophical investigation. This is the challenge taken up in CU. RQ2 concerns explainability. How effective are different approaches at rendering AI systems explainable? Backchaining through the model from *understanding* we reach *explanatory information*, *explainability approach*, and *AI system*, the constructs which pertain to this question. Given a

⁴³ In reality, this story is further complicated by considering that developers rely on pre-existing research during model development. In this specific example, relevant research areas would include uncertainty quantification in AI (Abdar et al., 2021), data visualisation (Franconeri et al., 2021; Pu & Kay, 2018), and reliance in collaborative (human-machine) decision-making (Guo et al., 2024; Schoeffer et al., 2024). The availability of high-quality and actionable research in all of these areas can ultimately affect desiderata satisfaction.

pragmatic definition of understanding, we need to know what kind of explanatory information can be produced about different AI systems by different explainability approaches (see Buchholz, 2023 for a related approach). This requires a careful appraisal of extant XAI approaches. XAI-MI and RPET serve this function. RQ3 concerns concrete manifestations of the BBP. Here, my goal is to apply the model to particular domains. While the entire model is relevant to this pursuit, it is particularly important to identify the relevant stakeholders, desiderata, and human-system interactions involved.⁴⁴ Then, one must evaluate *whether* opacity is preventing the satisfaction of those desiderata. If it is, one must consider whether and how explainability approaches can produce explanatory information that increases understanding (for the right stakeholder) and, consequently, enables desiderata to be satisfied. I do this in both DFB (for the domain of neuroscience) and UB (for data-driven psychiatry).

1.4 My Contributions

With the model of BBP contexts above in place, I now turn to a more detailed discussion of my dissertation contributions. My goals with this discussion are twofold. First, to explain how each paper addresses one of my three research questions. And second, to show how the articles fit into a shared conceptual framework. In what follows, I elaborate on each article one-by-one, with the articles grouped by the research question they address. I structure my discussion of each article the same way. I begin by contextualising each article (in its research literature). I then provide a summary of the content of the article and explain how it addresses one of my research questions. Finally, I clarify how the article fits within my model of the BBP.

⁴⁴ Naturally, each domain may have many such interactions worth considering. The relation between domains and BBP manifestations is one-to-many.

1.4.1 RQ1: How should *understanding* be construed in the context of the BBP?

Chapter 2: *Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach (CU)*

Research Context CU focuses on RQ1, the conceptual question of how we should construe understanding in the context of the BBP. Clearly, understanding plays a central role in the BBP as I have presented it. If generating and communicating explanatory information about AI systems is to benefit stakeholders, the route by which this must occur is through increasing (somebody's) understanding. This raises a problem for XAI research. Namely, that the concept of understanding is nebulous and ill-defined. Since, according to my model of the BBP, *understanding* is the epistemic construct that explanatory approaches should target, it is crucial for that construct to have a clear and practical interpretation. This state of affairs motivates CU.

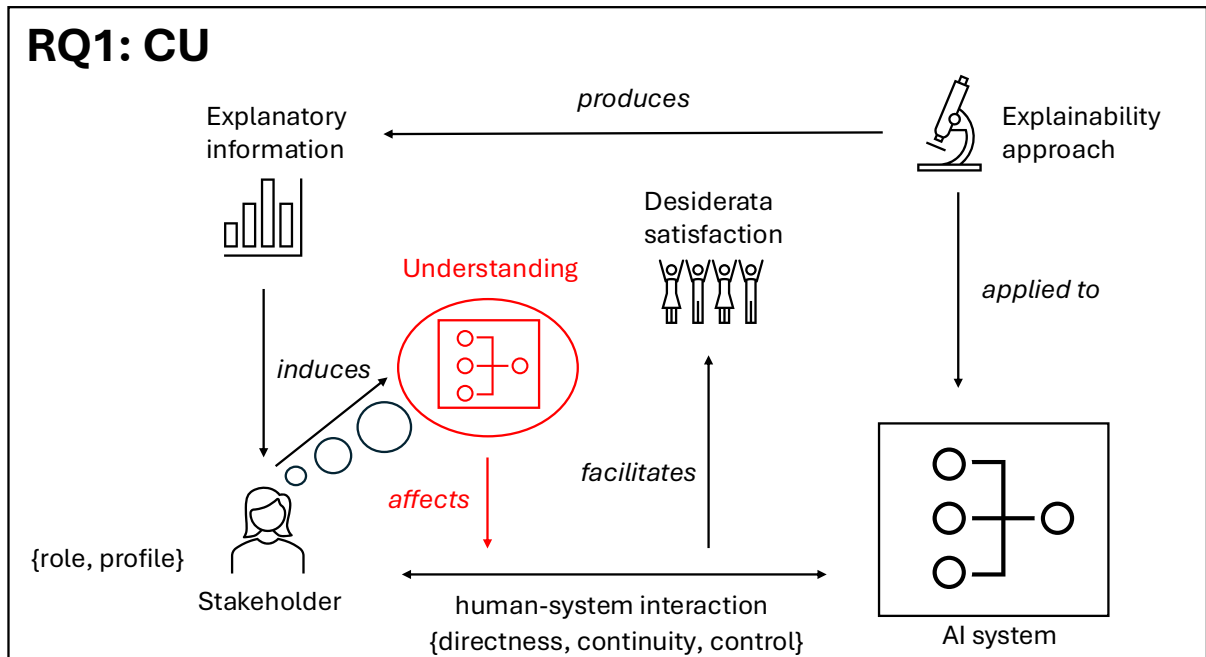
Summary In CU, my co-authors and I defend a pragmatic approach to *understanding* in the context of the BBP. First, we delineate criteria for a conceptualisation of understanding fitting for analysing BBP manifestations.⁴⁵ Then, we argue that an abilities-based approach, in which understanding is construed as possession of the right kind of abilities, satisfies those criteria. In particular, we argue that conceptualising understanding via abilities is i) theoretically motivated by convergent research across relevant disciplines, ii) versatile, in the sense that various abilities can be specified to capture different forms and degrees of understanding, iii) operationalisable, i.e., amenable to quantitative empirical investigation, and iv) related to desiderata, in the sense that the epistemic requirements for desiderata satisfaction in BBP manifestations can be spelled out concretely in terms of the abilities stakeholders need to be endowed with. To show that an abilities-based account satisfies the first criterion, we point to prominent research in philosophy, cognitive and educational psychology, causal inference, and human factors, all of which view abilities as central to understanding (see e.g., de Regt, 2009; Hills, 2016; Lombrozo & Carey, 2006; Pearl & Mackenzie, 2020). To address the versatility criterion, we distinguish six clusters of abilities – recognising,

⁴⁵ In CU we use the term *XAI contexts* to refer to what I am calling BBP manifestations here.

assessing, predicting, explaining, intervening, and designing – which capture the different forms of understanding relevant for BBP manifestations. To demonstrate the operationalisability of an abilities-based account of understanding, we point to prior work in XAI that has employed abilities to assess the efficacy of XAI methods (e.g., Piltaver et al., 2016; Poursabzi-Sangdeh et al., 2021; Ribeiro et al., 2016). Finally, to argue that abilities are suitable for mediating the relationship between explainability and desiderata satisfaction, we show how abilities can be used to specify epistemic requirements for the satisfaction of common desiderata such as fairness and informed self-advocacy.

Research Contribution To address RQ1, CU provides an account of understanding that balances the competing constraints of theoretical motivation and practical utility. As discussed in Chapter 1.2.2, there are numerous philosophical accounts of *understanding* in the literature (see Baumberger et al., 2016 for an overview). However, many of these accounts, whatever their theoretical virtues, fail to meet our standards for versatility and operationalisability. For example, accounts which describe understanding as the grasp of a correct explanation may be appropriate for scientific contexts (Strevens, 2013). But such accounts lack the versatility to capture how delivering partial and idealised explanatory information in formats tailored to stakeholders might aid them in navigating BBP manifestations (Fleisher, 2022; Keil, 2019). Providing a varied set of particular abilities which indicate the kind of understanding relevant for BBP manifestations helps solve this problem. On the other hand, however, ignoring theoretical considerations completely would risk severing the connection between our conceptualisation of understanding and the way the term is used in other academic contexts. To avoid this, we surveyed interdisciplinary literature on understanding, ensuring that our account is precededented across scientific fields which focus on the analysis and measurement of epistemic outcomes.

FIGURE 3: THE RESEARCH FOCUS OF CU IN CONTEXT



Fit to BBP Model In terms of the BBP model presented in Chapter 1.3.2, CU provides an interpretation of the central construct of *understanding* (see Fig. 3). The criteria which we use to evaluate our conceptualisation of understanding ensure that it is adapted to play its role in the model. In particular, *versatility* ensures that our conceptualisation can be applied to stakeholders with variable profiles (and epistemic requirements). *Operationalisability* ensures that the causal path between *explanatory information* and *understanding* can be investigated empirically, such that explanatory approaches can be rigorously assessed for their potential to bring about understanding. Finally, satisfying the *relation to desiderata* criterion ensures that our notion of understanding supports the *facilitation* relation that holds between *understanding* and *desiderata satisfaction*.

1.4.2 RQ2: How effective are extant explainability approaches?

Chapter 3: Explaining AI Through Mechanistic Interpretability (XAI-MI)

Research Context The next two chapters in my dissertation are concerned with RQ2, the explainability of AI systems. The XAI literature contains numerous surveys of explainability methods and attempts to model the BBP (Belle & Papantonis, 2021; Buchholz, 2023; Fleisher, 2022; Guidotti et al., 2018; Hoffman et al., 2018; Langer et al., 2021). One important function of these contributions is to portray the *landscape* of XAI, characterising extant problems and enumerating the options available for solving them

(Longo et al., 2024; Speith, 2022). These efforts tend to place a strong focus on the method-centric approach (see Table 2, Chapter 1.3.1). However, recall from the discussion in Chapter 1.3.1 that the deployment of XAI methods to deliver explanatory information to stakeholders *in situ* (that is, in the course of direct system-computer interaction) is just one of several XAI approaches. In my view, a narrow focus on the potential of method-centric approaches presents an impoverished view of the landscape of XAI research, failing to account for alternative approaches such as MI. The nature and potential of this approach is the topic of XAI-MI.

Summary In XAI-MI, my co-author and I describe MI and evaluate its prospects for ameliorating the opacity of AI systems. Whereas the field of XAI has primarily focussed on developing specific methods to produce *in situ* explanatory information about AI systems, MI applies coordinated discovery strategies to uncover the functional organisation of trained AI systems. In XAI-MI, we provide a philosophical analysis of MI as a *pattern discovery process* (Bechtel & Richardson, 2010; Haueis & Kästner, 2022; Kästner & Haueis, 2021). As such a process, MI involves the functional decomposition of target systems (i.e., ANNs), the localisation of subfunctions to discrete components within those systems (e.g., layers, neurons, circuits etc...), and the iterative refinement of functional hypotheses through systematic experimental manipulations (e.g, visualisations, ablations etc...). Working through the example of Distill’s investigation into curve circuits in the early layers of the ANN InceptionV1 (Cammarata et al., 2021), we demonstrate that MI features all the hallmarks of a pattern discovery process. As well as analysing MI philosophically, we also clarify how it differs from the method-centric approach to XAI⁴⁶ and discuss its distinctive costs and benefits. We argue that MI *can* successfully uncover the functional organisation of trained ANNs, producing detailed mechanistic explanations that provide leverage for improved prediction and control of opaque AI systems. However, we also acknowledge that the benefits of producing detailed mechanistic explanations come with large costs in terms of time, labour, and expertise. We claim that, given these features, MI is an important and distinctive

⁴⁶ We refer to this approach as the *divide-and-conquer* approach in XAI-MI.

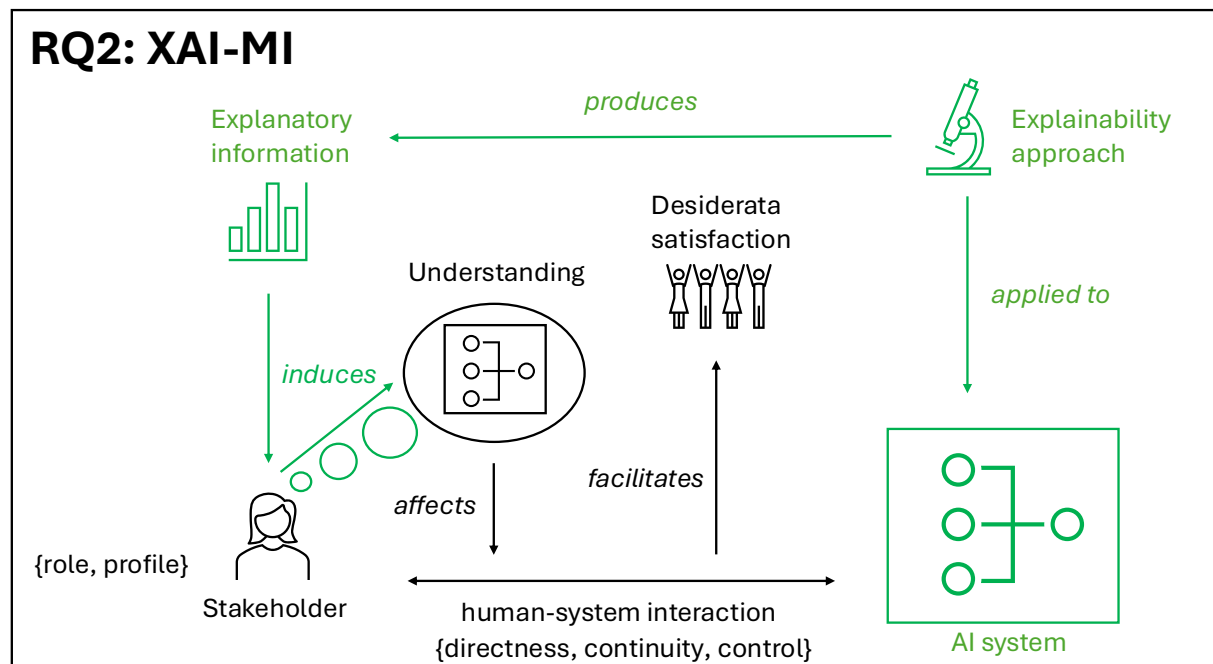
approach to XAI that ought to be applied when the benefits of detailed explanations are sufficient to render the costs worth bearing.

Research Contribution With respect to RQ2, one of the core goals of XAI-MI is to push back against the notion that the explainability of AI systems, viz., how well people *can* understand them, is well-approximated by what particular XAI methods can show about them. This view underestimates XAI, failing to appreciate the explanatory potential of hypothesis-driven discovery processes that iteratively refine mechanistic models of target systems. This mistake can lead to erroneous claims about the explainability of AI systems. For example, Rudin states that XAI, in general, fails to provide detailed explanatory information about target systems (Rudin, 2019; see Cearns et al., 2019; Fleisher, 2022; Freiesleben, 2024; Z. C. Lipton, 2018 for similar claims). This claim is plausible if XAI is considered as a collection of methods, but false if XAI is considered as a set of disparate explanatory approaches. Importantly, MI focuses on *expert* understanding.⁴⁷ Since what makes the BBP for AI systems unique is that the developers of the systems (viz., experts) do not understand how they work, careful consideration of the extent to which expert understanding can be facilitated is particularly crucial for understanding the contours of the BBP.

Fit to BBP Model In terms of the BBP model (see Fig. 4), XAI-MI investigates one particular *explainability approach*, showing that it can produce a particularly comprehensive and detailed kind of *explanatory information*. Further, the case study we present shows that this explanatory approach is effective in increasing *understanding*. In particular, one researcher was able to perform a computational recomposition of the curve circuit (i.e., setting the weights of an ANN to implement the circuit by hand, see XAI-MI for details). This demonstrates the ability *to design*, the most demanding ability detailed in CU (i.e., the ability indicating the greatest degree of understanding). At the time of writing, I am not aware of another XAI approach capable of inducing the ability to design a functioning AI system-component of comparable scale (see Geiger et al., 2021 for a methodology apt

⁴⁷ This is not unique to MI, the same is true of the theoretical approach to XAI.

FIGURE 4: THE RESEARCH FOCUS OF XAI-MI IN CONTEXT



for assessing this ability).⁴⁸ Thus, the abilities-based approach of CU can be used to express the distinctive strengths of MI, reinforcing our claims about the versatility of our abilities *and* the potential of MI. XAI-MI also illustrates the need to conceive of BBP manifestations broadly. Though we argue that MI can support the satisfaction of desiderata such as safety and trustworthiness, this is an *indirect* process. MI produces explanatory information that contributes to a store of knowledge about trained ANNs' functional organisation. This store can influence the design of future systems, ensuring that they avoid known vulnerabilities (e.g., to adversarial attacks). Thus, the application of explainability approaches to AI systems can be temporally distant (i.e., years apart) from the direct stakeholder-system interactions in which desiderata are satisfied.⁴⁹

Chapter 4: Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)

Research Context The next chapter in my dissertation continues to address RQ2, focusing on the topic of performance-explainability trade-offs (PETs). The strong

⁴⁸ The entire ante-hoc explainability approach involves treating the ability to design as a strict requirement. However, an important flaw of this approach is that demanding this ability risks restricting developers to uncompetitive systems (see the discussion of RPET).

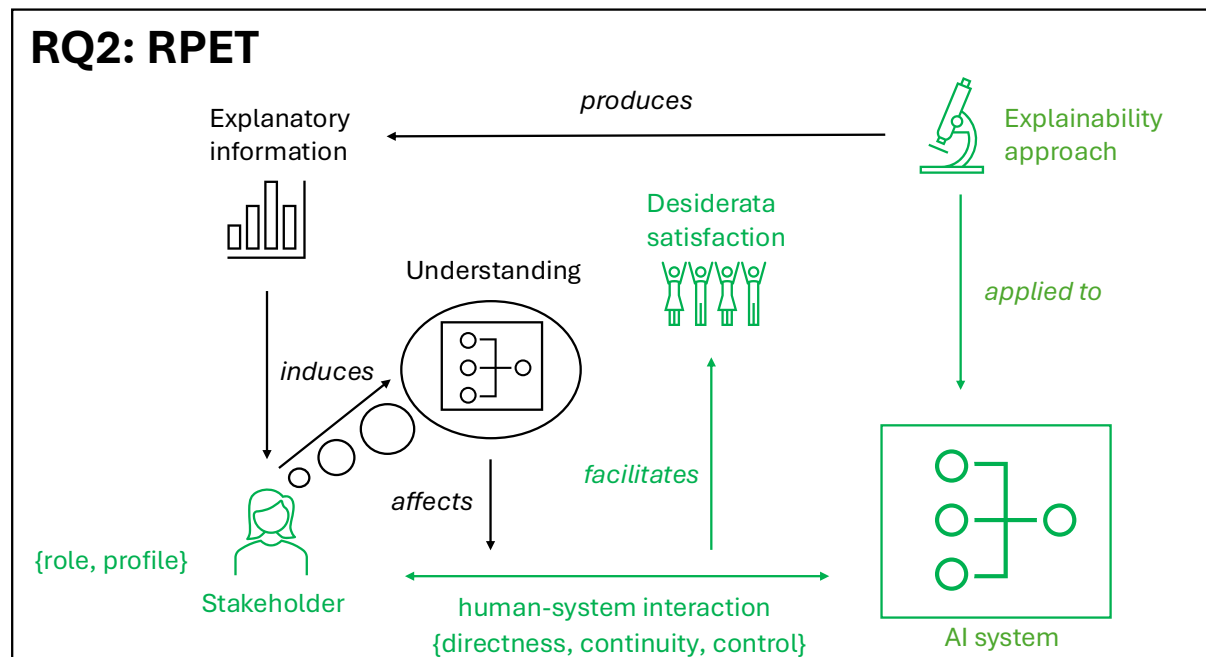
⁴⁹ Note that this is not at all unusual. We routinely use technological products whose safety is ensured by guarantees that certain protocols were followed years or decades ago.

performance of opaque AI models across domains (see Chapter 1.2.1) prompted claims that developers and deployers face trade-offs when constructing and selecting models. The basic case is as follows. If, among the set of models one might develop or deploy for a given task, those which achieve the best performance are opaque, and at least one of those which achieves lower performance is not, then the choice of which model to use for that task involves a trade-off between the values of explainability and performance. This is entailed by the fact that no choice will simultaneously maximise both values (Matthewson & Weisberg, 2009; Tversky & Kahneman, 1986). The claim that PETs obtain in real-world contexts has been made in the XAI literature (e.g., Barredo Arrieta et al., 2020; Gunning & Aha, 2019). In RPET, my co-authors and I critically analyse the case for a performance-explainability trade-off.

Summary In RPET, we begin by presenting the case for a trade-off, as described above. We then introduce criticism levied against that case, focusing on an influential paper by Cynthia Rudin (2019). Rudin claims 1) that the widespread existence of the PET is a myth, 2) that XAI has little hope of ameliorating the opacity of black box models, and, therefore, 3) that the use of ante-hoc explainable models is preferable in almost every real-world context. We argue that Rudin’s analysis of the PET stacks the deck in favour of her preferred conclusions. In particular, she affords greater resources to the workflow she is familiar with (iterative development of ante-hoc explainable models) and underestimates the potential of post-hoc XAI (equating it with the method-centric approach and ignoring the potential of MI). We use these counterarguments as a catalyst to present a framework for analysing PETs. Our framework casts the PET as a multi-objective optimisation problem that explicitly includes *developmental resources*, *domain features*, and *risks*. On our view, whether a PET obtains is a question best applied to rival model development procedures (viz., trajectories of actions developers can take), as opposed to specific models or model classes. Crucially, a fair analysis must hold fixed the quantity of resources distinct procedures have access to. Further, the performance and explainability values attainable for different model development choices depend on features of the data domain being modelled. And the relative values of performance and explainability for a given context are conditional on the kind of risks which obtain in that context (e.g., sometimes explainability is worth sacrificing for performance).

Research Contribution The contributions of RPET to RQ2 are twofold. First, our framework codifies the insight that concerns about PETs are best considered at the level of model-development procedures (not choices of model class). This allows comparisons between explainability approaches to be conducted in a fair and principled way. Second, we provide reasons to doubt Rudin’s more ambitious claims about the prospects of ante-hoc explainability. Notably, our analysis posits that the notion of *explainability* most relevant to the BBP is not an objective property of models (or model classes), but a complicated function of model type, domain features, and development resources (Freitas, 2014; Huysmans et al., 2011). This serves as a corrective to *both* unwarranted assumptions that the PET is ubiquitous *and* Rudin’s suggestion that it rarely or never holds. Evidence strongly suggests that there *are* domains for which black-box approaches achieve greater performance than ante-hoc explainable ones (and thus that PETs do obtain).⁵⁰ However, for many other domains of interest, the jury remains out.⁵¹

FIGURE 5: THE RESEARCH FOCUS OF RPET IN CONTEXT



⁵⁰ For one prominent example consider the case of AlphaFold2 (Jumper et al., 2021), a data-intensive DL-based approach to predicting proteins’ three-dimensional folded structure from their amino acid sequences. In this domain, the iterative development of ante-hoc explainable models guided by expert domain knowledge was pursued for decades before being decisively outperformed by a black-box approach (Bouatta et al., 2021). The notion that no PET obtains in such cases is wildly implausible.

⁵¹ Following Rudin, we take the question of whether data domains contain *hidden patterns* to be crucial (see Chapter 4.3.2). Hidden patterns are complex, non-linear, and/or distributed features that capture (robustly) predictive information in the input data (see also Boge, 2022, 2023 for related discussion).

Fit to BBP Model The analysis presented in RPET can be expressed in terms of the BBP model (see Fig. 5). As discussed above, analysis of PETs is most sensibly applied to model development procedures. In the context of the BBP model, distinct procedures involve different combinations of AI systems (e.g., decision tree/ANN) and explainability approaches (e.g., ante-hoc/MI). Performance and explainability are the desiderata under consideration (which, according to proponents of the PET, sometimes conflict). In RPET's framework, developer expertise is treated as a resource which determines the range of actions the developer can take (e.g., how effectively they can apply an MI approach). In the BBP model, this would be captured by the stakeholder profile. For fair comparison of model development procedures, further developmental factors (e.g., time, money) and details of the data domain (e.g., complexity) should be held fixed and are thus (crucial) background information rather than variables in the model.

1.4.3 RQ3: To what extent can specific manifestations of the BBP be solved?

Chapter 5: Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience

Research Context With the abilities-based notion of understanding and a nuanced view of the explainability of AI systems in mind, we now turn to RQ3: concrete BBP manifestations. Neuroscience is one domain in which it has been argued that the opacity of AI systems limits desiderata satisfaction (Bowers et al., 2022; Chirimuuta, 2021). In neuroscience, deep learning is an increasingly popular method for modelling perceptual and cognitive neural systems (Cichy & Kaiser, 2019; Richards et al., 2019). In this research program, neuroscientists develop ANNs to implement cognitive tasks that living organisms are hypothesised to perform, such as object recognition (Yamins & DiCarlo, 2016; Zhuang et al., 2021). They then evaluate the quality of these models with respect to three key properties (Cao & Yamins, 2021; Doerig et al., 2023; Schrimpf et al., 2020). First, can the ANN implement the task (i.e., achieve performance comparable to the target system)? Second, does the ANN display behavioural similarity to the target system (e.g., succeed and err in similar cases)? Third, does the ANN display neural similarity to the target system (e.g., are neural activations in subparts of the ANN predictive of activations in the subparts of the target system)? When ANNs perform well according to these criteria, proponents of the research program take this as justification for

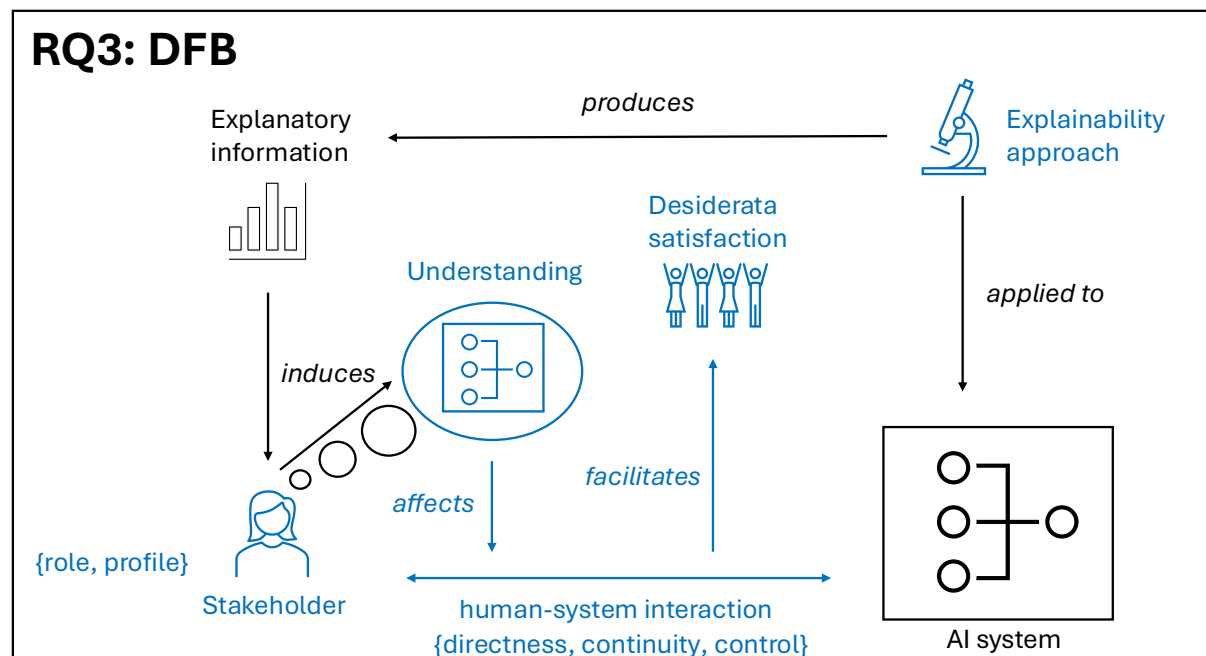
considering them successful scientific models of their explanatory targets (Bowers et al., 2022). However, critics of this approach worry that the predictive adequacy displayed by ANN models fails to produce understanding of target systems (Chirimuuta, 2021). Whether such worries are well founded is the topic addressed in DFB.

Summary In DFB, we reject claims that pursuing a deep learning-driven research agenda in neuroscience entails foregoing understanding. Instead, we claim, ANN-based modelling of cognitive and perceptual systems offers a novel epistemic perspective that complements traditional approaches. Critics of the deep learning approach invoke the same worries that motivate the general BBP. Namely, they are concerned that ANN models of biological neural systems achieve predictive accuracy but fail to offer human-interpretable insights into how those systems work (Chirimuuta, 2021; Lindsay, 2021). The issue is that neuroscientists can develop predictively adequate ANN models without understanding either 1) how they implement their behaviour, or 2) how that implementation relates to the functioning of the target system (i.e., the biological neural network) (this is what Sullivan, 2022 calls ‘link uncertainty’). Since one could pursue the development of alternative models whose functioning is easier to grasp (e.g., Georgopoulos et al., 1986), it is intuitively plausible that the deep learning approach trades off understanding for predictive accuracy. However, in DFB, we show that this argument suffers from a subtle conceptual mistake. How easy it is to understand a model cannot be equated with the degree of understanding one might gain through protracted interaction with that model. In line with CU’s abilities-based approach, we invoke a notion of understanding based around what people are enabled to *do* (e.g., predict, intervene) via development of and interaction with models (see also Lawler & Sullivan, 2021). This involves in-depth post-hoc explainability (e.g., MI) and integrating an interpretation of model properties with complementary sources of information (e.g., neurobiological data) (Willeke et al., 2023). With this view of understanding in mind, the deep learning-based approach has a realistic route to inducing neuroscientific understanding.

Research Contribution RQ3 concerns the solubility of specific manifestations of the BBP. Philosophers of science from several disciplines have expressed concern about the increasing use of opaque ANNs to model target phenomena (Boge, 2022; Boon, 2020;

Lindsay, 2021; López-Rubio & Ratti, 2021; Srećković et al., 2022). These concerns are not unfounded, but the suggestion that deep learning approaches *necessarily* invite a trade-off requires careful scrutiny. In particular, the mere observation that ANNs are opaque, in the sense that they exhibit significant barriers to human understanding, is not sufficient justification for that claim. In line with this, an important contribution of DFB is presenting a more detailed view of the dependency structure on which such claims rely. In particular, while models play a central role in neuroscience, the abilities which indicate understanding are typically supported by a distributed and heterogeneous web of explanatory information (Burnston, 2016; Hochstein, 2016). Models are not interpreted in a vacuum. Thus, a realistic assessment of what a particular research direction can contribute to scientific understanding ought to consider how the explanatory information produced (e.g., an ANN model and subsequent post-hoc analyses) interacts with the extant body of explanatory information. With this view in mind, the manifestation of the BBP for neuroscience is plausibly soluble. Further, while our claims in DFB are focused on neuroscience specifically, the philosophical approach can be applied to questions about the effects of deep learning in other scientific fields.

FIGURE 6: THE RESEARCH FOCUS OF DFB IN CONTEXT



Fit to BBP Model The argument we make in DFB follows from the application of the BBP model (see Fig. 6) to the neuroscience context.⁵² The key places where we depart from proponents of the trade-off view are our notion of understanding, our assessment of the potential of explainability approaches, and our construal of the human-system interaction. In terms of understanding, the pragmatic view that it should be equated with particular abilities (as per CU) leads away from equating understanding with the grasp of a particular model (Lawler & Sullivan, 2021). If understanding a target system is partially grounded in an ability to predict, explain, and control its behaviour (Lombrozo & Carey, 2006; Pearl & Mackenzie, 2020; Woodward, 2003), then various sources of information will support these feats. In terms of explainability approaches, the potential of MI to uncover the functional structure of trained ANNs is particularly relevant to cognitive neuroscience. It is the opacity of precisely this aspect of ANNs that leads to pessimistic arguments about the prospects of deep learning-driven approaches (Lindsay & Bau, 2023). In terms of human-system interactions, we diverge from pessimists in considering temporally extended interaction. We acknowledge that ANNs pose significant barriers to understanding, but scientific research is a domain in which protracted attempts to overcome opacity are commonplace and acceptable. Finally, consideration of stakeholder profiles also plays a role in this debate. The training of neuroscientists leaves them well-placed to apply post-hoc explainability approaches to ANN models (Ivanova et al., 2021; Lindsay & Bau, 2023; Willeke et al., 2023).

Chapter 6: *Understanding as a Bottleneck for the Data-driven Approach to Psychiatric Science*

Research Context The final paper in my PhD dissertation focuses on another concrete domain. In psychiatric science, data-driven approaches promise to improve clinical decision-making by incorporating large volumes of information from multiple modalities (e.g., survey data, biological data, behavioural data) (Bzdok & Meyer-Lindenberg, 2018; Fernandes et al., 2017; Gillan & Whelan, 2017). Making use of deep learning methods, ANNs can be employed to diagnose patients, offer prognoses, and assign treatments (Chekroud et al., 2021; Janssen et al., 2018; Sharma & Verbeke, 2020). On an optimistic

⁵² Note that the relation between *understanding* and *desiderata satisfaction* is trivial in this case because understanding is itself the desideratum we are considering.

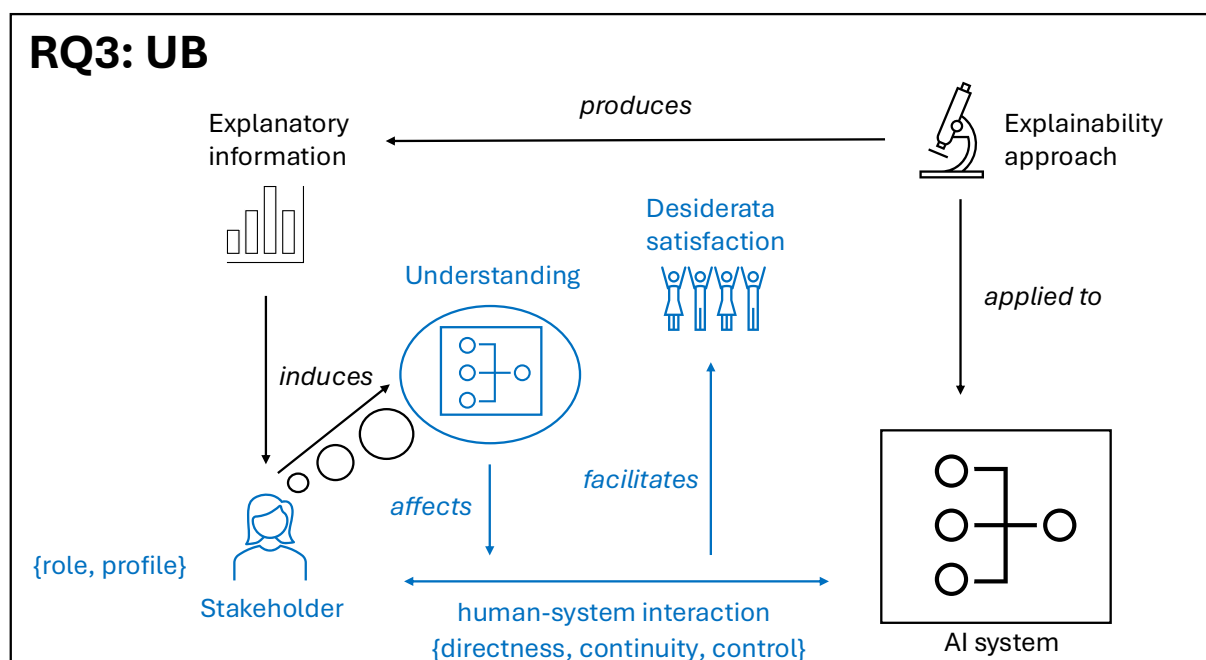
view, this data-driven approach will prove transformative for psychiatry, radically improving patient outcomes by assigning the right treatment at the right time (Fernandes et al., 2017). In UB, I critically assess this claim.

Summary In UB, I present the data-driven approach to psychiatry as one grounded in a view of clinical decisions (diagnosis, prognosis, treatment assignment) as functions (i.e., input-output mappings). Since there is compelling evidence that mental illnesses are massively multifactorial (Kendler, 2016; Kendler & Gynge, 2020), using large volumes of data to develop complex models approximating these functions seems a promising avenue of research. Proponents of the data-driven approach hope that relying on data can free psychiatry from the shackles of theoretical commitments (e.g., about the nature and structure of mental illness) (Chekroud et al., 2021; Shickel et al., 2018). However, though data-driven approaches are valuable, in UB, I argue that identifiable pathways by which the data-driven approach to psychiatry might bring about transformative improvements to mental health outcomes are blocked by *patient understanding*, viz., the understanding a patient has of their own condition. I construe this form of understanding as the possession of *adaptive mental constructs* through which experience is mediated (following Dilthey, 1894/1977). I contend that the data-driven approach has no clear means of addressing patient understanding. The limitations of extant treatment options bound the possible gains from improving allocation via improved diagnosis and treatment assignment. And the development of better treatments cannot be effectively pursued in a theory-free manner. Finally, though applying explainability approaches to complex models of psychiatric illness may provide insights into their nature, leveraging these insights to improve patient outcomes is an unsolved problem. Practitioners may lack the ability to explain scientific insights to patients in a way which increases desiderata satisfaction. Or, to frame it another way, patients may lack the ability to integrate these insights into their “assumptive worlds” (Frank & Frank, 1993, pp. 37–38).

Research Contribution UB delivers several contributions to RQ3 and the broader philosophical literature. Starting with RQ3, my argument denies that successful explanation of opaque AI systems would enable the transformative effects the data-driven approach hopes for. This brings into question whether the opacity of the AI systems is what prevents desiderata satisfaction in this case, and thus whether it is a BBP

manifestation (see also Sheu, 2020). I take this as a valuable demonstration that the involvement of black-box AI systems in a domain do not automatically produce BBP manifestations. Beyond the scope of RQ3, UB makes several further contributions. First, optimism about the data-driven approach is often unaccompanied by a detailed description of how exactly it will bring about transformative change. Focussing on patient understanding forces one to consider concrete paths to transformativity and thus reveals the assumptions that the success of the data-driven approach depends on (see also Horn & Weisz, 2020). Second, the philosophy underlying the data-driven approach to psychiatry is underspecified (though see Bickman, 2020).⁵³ Making the function-oriented, theory-neutral philosophy of the research paradigm explicit aids in assessing its potential. Finally, in presenting an account of understanding as *adaptive mental constructs*, UB synthesises accounts of understanding from the classic and recent philosophical literature on psychiatry and modern philosophy of science (Dilthey, 1894/1977; Wampold et al., 2007; Wilkenfeld, 2013).

FIGURE 7: THE RESEARCH FOCUS OF UB IN CONTEXT



Fit to BBP Model Although I am sceptical that the problems facing the data-driven approach to psychiatry constitute a manifestation of the BBP, my analysis can still be

⁵³ This may be due to the technical challenges associated with the research, the aversion to philosophy captured by the ideal of theory-neutrality, or a combination of the two.

usefully framed in terms of the BBP model (see Fig. 7). Psychiatry is a domain in which the profiles of the relevant stakeholders and the indirectness of the human-system interaction are particularly important. Model developers are technically proficient practitioners who apply their expertise to produce powerful predictive models of psychiatric disorders (Dwyer et al., 2018; Koutsouleris et al., 2021; Rutledge et al., 2019). Psychiatrists and other mental health professionals often have a moderate understanding of AI and hope to employ it to ameliorate the suffering of their patients (C. Brown et al., 2021; Horn & Weisz, 2020). And psychiatric patients themselves typically have minimal background knowledge about AI models and how they might aid a process of coping and recovery. Further, they often interact with clinical support systems only indirectly (i.e., via a medical practitioner) and with limited control. These considerations are crucial to my argument. Optimists might be right to think that ANNs and other complex models will capture important underlying structure in mental illnesses and psychopathology (Shickel et al., 2018).⁵⁴ And, via explainability approaches, technical researchers may be able to uncover, describe, and understand that structure (pace Cearns et al., 2019). However, even if these conditions hold, it is not at all clear that the kind of explanatory information produced would be deliverable to patients in a form that would induce their understanding, which, according to UB, is required for improved outcomes (Frank & Frank, 1993; Wampold, 2015).

1.5 Summary

As Alan Turing foresaw in 1950, machines that learn from data may be opaque to their developers (Turing, 1950, p. 459). Today, as AI systems developed with machine learning techniques become increasingly powerful, their opacity is generating complex and multifaceted problems. To assess the solubility of these black box problems, the careful integration of conceptual and empirical insights from a variety of academic disciplines is necessary. In this dissertation, I approach the BBP in a piecemeal manner, carving off

⁵⁴ Note that this is a recurrence of the same issue discussed in RPET (Chapter 4.3.2) under the label of *hidden patterns*.

semi-independent questions which can be tackled within the strictures of independent articles. This introduction presents a common framework for these research contributions, showing how they are situated in a unified model of the BBP.

To conclude the introduction, I revisit my three research questions and summarise how my work addresses them. RQ1 concerns the notion of *understanding* most apt for BBP manifestations. This is directly addressed by the abilities-based account proposed in CU. RQ2 considers how effectively extant explainability approaches produce explanatory information that induces understanding and increases desiderata satisfaction. XAI-MI contributes to this question by characterising the mechanistic interpretability approach in a way that highlights its distinctive cost-benefit profile. And RPET clarifies how we should reason about and compare the value of different explainability approaches. RQ3 investigates the solubility of the BBP in concrete contexts. The argument presented in DFB suggests that the use of opaque ANN models *can* contribute to increased neuroscientific understanding, countering concerns raised by critics. UB's analysis of the data-driven approach to psychiatry, however, provides a more pessimistic view, arguing that the approach has no clear means of targeting patient understanding. Overall, these research contributions shed light on how we should think about black box problems and how we might overcome them.

1.6 References

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297. <https://doi.org/10.1016/j.inffus.2021.05.008>
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2024). Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy*, 33(2), 327–337. <https://doi.org/10.1111/jems.12521>
- Alvarez, M. (2009). How many kinds of reasons? *Philosophical Explorations*, 12(2), 181–193. <https://doi.org/10.1080/13869790902838514>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. <https://doi.org/10.48550/arXiv.1606.06565>

- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Andersen, H. (2018). Complements, Not Competitors: Causal and Mathematical Explanations. *The British Journal for the Philosophy of Science*, 69(2), 485–508. <https://doi.org/10.1093/bjps/axw023>
- Arbib, M. (1961). Turing Machines, Finite Automata and Neural Nets. *Journal of the ACM*, 8(4), 467–475. <https://doi.org/10.1145/321088.321089>
- Artificial Intelligence Act (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf
- Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., ... Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624), 1067–1074. <https://doi.org/10.1126/science.ade9097>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bastani, O., Kim, C., & Bastani, H. (2019). *Interpreting Blackbox Models via Model Extraction* (arXiv:1705.08504). arXiv. <https://doi.org/10.48550/arXiv.1705.08504>
- Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, 35(1), 12. <https://doi.org/10.1007/s13347-022-00510-w>
- Baumberger, C. (2019). Explicating Objectual Understanding: Taking Degrees Seriously. *Journal for General Philosophy of Science*, 50(3), 367–388. <https://doi.org/10.1007/s10838-019-09474-6>
- Baumberger, C., Beisbart, C., & Brun, G. (2016). What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science. In *Explaining Understanding*. Routledge.
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–441. <https://doi.org/10.1016/j.shpsc.2005.03.010>
- Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. The MIT Press; JSTOR. <https://doi.org/10.2307/j.ctt5hhhpX>
- Beisbart, C. (2021). Opacity thought through: On the intransparency of computer simulations. *Synthese*, 199(3), 11643–11666. <https://doi.org/10.1007/s11229-021-03305-2>
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National*

Academy of Sciences, 116(32), 15849–15854.

<https://doi.org/10.1073/pnas.1903070116>

Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4.

<https://www.frontiersin.org/articles/10.3389/fdata.2021.688969>

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. <https://doi.org/10.1109/tpami.2013.50>

Bickman, L. (2020). Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. *Administration and Policy in Mental Health*, 47(5), 795–843.

<https://doi.org/10.1007/s10488-020-01065-8>

Boddington, P. (2023). Normative Ethical Theory and AI Ethics. In P. Boddington (Ed.), *AI Ethics: A Textbook* (pp. 229–276). Springer Nature. https://doi.org/10.1007/978-981-19-9382-4_6

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1), 43–75. <https://doi.org/10.1007/s11023-021-09569-4>

Boge, F. J. (2023). Functional Concept Proxies and the Actually Smart Hans Problem: What's Special About Deep Neural Networks in Science. *Synthese*, 203(1), 16.

<https://doi.org/10.1007/s11229-023-04440-8>

Boon, M. (2020). How Scientists Are Brought Back into Science—The Error of Empiricism. *A Critical Reflection on Automated Science: Will Science Remain Human?*, 43–65. https://doi.org/10.1007/978-3-030-25001-0_4

Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section D, Structural Biology*, 77(Pt 8), 982–991.

<https://doi.org/10.1107/S2059798321007531>

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolphi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *The Behavioral and Brain Sciences*, 1–74. <https://doi.org/10.1017/S0140525X22002813>

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

Bringsjord, S., & Govindarajulu, N. S. (2024). Artificial Intelligence. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024). Metaphysics Research Lab, Stanford University.

<https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/>

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [Cs, Stat]*.

<http://arxiv.org/abs/2104.13478>

- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1), 139–159. [https://doi.org/10.1016/0004-3702\(91\)90053-M](https://doi.org/10.1016/0004-3702(91)90053-M)
- Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry*, 218(3), 131–134. <https://doi.org/10.1192/bjp.2019.245>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodi, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>
- Buchholz, O. (2023). A Means-End Account of Explainable Artificial Intelligence. *Synthese*, 202(2), 33. <https://doi.org/10.1007/s11229-023-04260-w>
- Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, 195(12), 5339–5372. <https://doi.org/10.1007/s11229-018-01949-1>
- Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), e12625. <https://doi.org/10.1111/phc3.12625>
- Burnston, D. C. (2016). Data graphs and mechanistic explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 57, 1–12. <https://doi.org/10.1016/j.shpsc.2016.01.002>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512. <https://doi.org/10.1177/2053951715622512>
- Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3. <https://doi.org/10.1016/j.bpsc.2017.11.007>
- Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve Circuits. *Distill*, 6(1), e00024.006. <https://doi.org/10.23915/distill.00024.006>
- Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. *arXiv:2104.01489 [Cs, q-Bio]*. <http://arxiv.org/abs/2104.01489>
- Carruthers, P. (1987). Conceptual pragmatism. *Synthese*, 73(2), 205–224. <https://doi.org/10.1007/BF00484740>
- Cartwright, N. (1994). *Nature’s Capacities and Their Measurement* (1st ed.). Oxford University Press/Oxford. <https://doi.org/10.1093/0198235070.001.0001>
- Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, 9(1), Article 1. <https://doi.org/10.1038/s41398-019-0607-2>
- Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*:

Official Journal of the World Psychiatric Association (WPA), 20(2), 154–170.

<https://doi.org/10.1002/wps.20882>

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings.

Computers in Human Behavior, 137, 107406.

<https://doi.org/10.1016/j.chb.2022.107406>

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199(1), 767–790. <https://doi.org/10.1007/s11229-020-02713-0>

Chollet, F. (2019). *On the Measure of Intelligence* (arXiv:1911.01547). arXiv.

<https://doi.org/10.48550/arXiv.1911.01547>

Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Simon and Schuster.

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4), 305–317. <https://doi.org/10.1016/j.tics.2019.01.009>

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3), 355–376.

<https://doi.org/10.1007/s11229-006-9097-x>

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.

Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4), 568–589. <https://doi.org/10.1086/709729>

de Regt, H. W. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, 76(5), 585–597. <https://doi.org/10.1086/605795>

de Regt, H. W. (2019). From Explanation to Understanding: Normativity Lost? *Journal for General Philosophy of Science*, 50(3), 327–343. <https://doi.org/10.1007/s10838-019-09477-3>

de Regt, H. W., & Dieks, D. (2005). A Contextual Approach to Scientific Understanding. *Synthese*, 144(1), 137–170. <https://doi.org/10.1007/s11229-005-5000-4>

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, 3(6), 100489. <https://doi.org/10.1016/j.patter.2022.100489>

Deck, L., Schomäcker, A., Speith, T., Schöffner, J., Kästner, L., & Köhl, N. (2024). *Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle* (arXiv:2404.18736).

arXiv. <https://doi.org/10.48550/arXiv.2404.18736>

Dierckx, L., Veroneze, R., & Nijssen, S. (2023). RL-Net: Interpretable Rule Learning with Neural Networks. In H. Kashima, T. Ide, & W.-C. Peng (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 95–107). Springer Nature Switzerland.

https://doi.org/10.1007/978-3-031-33374-3_8

Dilthey, W. (1977). Ideas Concerning a Descriptive and Analytic Psychology (1894) (R. M. Zaner, Trans.). In W. Dilthey (Ed.), *Descriptive Psychology and Historical Understanding* (pp. 21–120). Springer Netherlands. https://doi.org/10.1007/978-94-009-9658-8_2 (Original work published 1894)

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The

- neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1–20. <https://doi.org/10.1038/s41583-023-00705-w>
- Doyle, Y., Egan, S., Graham, N., & Khalifa, K. (2019). Non-factive Understanding: A Statement and Defense. *Journal for General Philosophy of Science*, 50(3), 345–365. <https://doi.org/10.1007/s10838-019-09469-3>
- Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics-2020-106820. <https://doi.org/10.1136/medethics-2020-106820>
- Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, 24(11), Article 11. <https://doi.org/10.1038/s41380-019-0365-9>
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, 14, 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Eden, A. H. (2007). Three Paradigms of Computer Science. *Minds and Machines*, 17(2), 135–167. <https://doi.org/10.1007/s11023-007-9060-8>
- Eldan, R., & Shamir, O. (2016). The Power of Depth for Feedforward Neural Networks. *Conference on Learning Theory*, 907–940. <https://proceedings.mlr.press/v49/eldan16.html>
- Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, 132(1), 33–42. <https://doi.org/10.1007/s11098-006-9054-z>
- Elgin, C. (2017). *True enough*. MIT Press.
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., ... Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.
- Erasmus, A., Brunet, T. D. P., & Fisher, E. (2021). What is Interpretability? *Philosophy & Technology*, 34(4), 833–862. <https://doi.org/10.1007/s13347-020-00435-2>
- Facchini, A., & Termine, A. (2022). Towards a Taxonomy for the Opacity of AI Systems. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2021* (pp. 73–89). Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_7
- Faconti, G. P., & Massink, M. (2000). Continuity in human computer interaction. *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, 364. <https://doi.org/10.1145/633292.633511>
- Falcon, A. (2023). Aristotle on Causality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality/>
- Faye, J. (2014a). Not Just Why-questions. In J. Faye (Ed.), *The Nature of Scientific Thinking: On Interpretation, Explanation, and Understanding* (pp. 210–240). Palgrave Macmillan UK. https://doi.org/10.1057/9781137389831_9

- Faye, J. (2014b). The Pragmatics of Explanation. In J. Faye (Ed.), *The Nature of Scientific Thinking: On Interpretation, Explanation, and Understanding* (pp. 183–209). Palgrave Macmillan UK. https://doi.org/10.1057/9781137389831_8
- Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26(6), 3333–3361. <https://doi.org/10.1007/s11948-020-00276-4>
- Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of ‘precision psychiatry’. *BMC Medicine*, 15(1), 80. <https://doi.org/10.1186/s12916-017-0849-x>
- Firestone, C. (2020). Performance vs. Competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43), 26562–26571. <https://doi.org/10.1073/pnas.1905334117>
- Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, 19(4), 534–560. <https://doi.org/10.1017/epi.2022.39>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3), 110–161. <https://doi.org/10.1177/15291006211051956>
- Frank, J. D., & Frank, J. B. (1993). *Persuasion and Healing: A Comparative Study of Psychotherapy*. JHU Press.
- Frankle, J., & Carbin, M. J. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv*. <https://dspace.mit.edu/handle/1721.1/129953>
- Freiesleben, T. (2024). *Artificial Neural Nets and the Representation of Human Concepts* (arXiv:2312.05337). arXiv. <https://doi.org/10.48550/arXiv.2312.05337>
- Freitas, A. A. (2014). Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1), 1–10. <https://doi.org/10.1145/2594473.2594475>
- Garcez, A. d’Avila, & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11), 12387–12406. <https://doi.org/10.1007/s10462-023-10448-w>
- Geer, A. J. (2021). Learning earth system models from observations: Machine learning or data assimilation? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200089. <https://doi.org/10.1098/rsta.2020.0089>
- Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, 34, 9574–9586. <https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html>
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing Causal Structure for Interpretable Neural Networks. *Proceedings of the 39th*

International Conference on Machine Learning, 7324–7338.

<https://proceedings.mlr.press/v162/geiger22a.html>

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science (New York, N.Y.)*, 233(4771), 1416–1419.

<https://doi.org/10.1126/science.3749885>

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, 18, 34–42.

<https://doi.org/10.1016/j.cobeha.2017.07.003>

Glock, H.-J., & Schmidt, E. (2021). Pluralism about practical reasons and reason explanations. *Philosophical Explorations*, 24(2), 119–136.

<https://doi.org/10.1080/13869795.2021.1908578>

Goh, G., † N. C., † C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3), e30.

<https://doi.org/10.23915/distill.00030>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system. In *Explanation and cognition* (pp. 299–323). The MIT Press.

Grimm, S. R. (2006). Is Understanding A Species Of Knowledge? *The British Journal for the Philosophy of Science*. <https://doi.org/10.1093/bjps/axl015>

Grimm, S. R. (Ed.). (2019). *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*. Oxford University Press.

<https://doi.org/10.1093/oso/9780190860974.001.0001>

Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of Explainable Artificial Intelligence. In M. Sayed-Mouchaweh (Ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems* (pp. 9–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-76409-8_2

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 93:1-93:42. <https://doi.org/10.1145/3236009>

Gunning, D., & Aha, D. (2019). DARPA’s Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2), Article 2. <https://doi.org/10.1609/aimag.v40i2.2850>

Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). A Decision Theoretic Framework for Measuring AI Reliance. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 221–236. <https://doi.org/10.1145/3630106.3658901>

Gurova, L. (2022). The Uses of Truth: Is There Room for Reconciliation of Factivist and Non-Factivist Accounts of Scientific Understanding? *International Studies in the Philosophy of Science*, 35(3–4), 211–221.

<https://doi.org/10.1080/02698595.2022.2084329>

- Haueis, P., & Kästner, L. (2022). Mechanistic inquiry and scientific pursuit: The case of visual processing. *Studies in History and Philosophy of Science*, 93, 123–135. <https://doi.org/10.1016/j.shpsa.2022.03.007>
- Hempel, C. (1965). Aspects of Scientific Explanation. In *Philosophy and Phenomenological Research* (p. 504). Free Press.
- Hempel, C., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15, 135–175.
- Hey, T., Butler, K., Jackson, S., & Thiyagalingam, J. (2020). Machine learning and big scientific data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2166), 20190054. <https://doi.org/10.1098/rsta.2019.0054>
- Hieronymi, P. (2011). XIV—Reasons for Action. *Proceedings of the Aristotelian Society*, 111(3_pt_3), 407–427. <https://doi.org/10.1111/j.1467-9264.2011.00316.x>
- Hills, A. (2016). Understanding Why. *Noûs*, 50(4), 661–688. <https://doi.org/10.1111/nous.12092>
- Hitchcock, C. R. (1995). Discussion: Salmon on Explanatory Relevance. *Philosophy of Science*, 62(2), 304–320. <https://doi.org/10.1086/289858>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, 193(5), 1387–1407. <https://doi.org/10.1007/s11229-015-0844-8>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018, December 11). *Metrics for Explainable AI: Challenges and Prospects*. arXiv.Org. <https://arxiv.org/abs/1812.04608v2>
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2024). Training compute-optimal large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030.
- Horn, R. L., & Weisz, J. R. (2020). Can Artificial Intelligence Improve Psychotherapy Research and Practice. *Administration and Policy in Mental Health*, 47(5), 852–855. <https://doi.org/10.1007/s10488-020-01056-9>
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
- Hullman, J., Holtzman, A., & Gelman, A. (2023). *Artificial Intelligence and Aesthetic Judgment* (arXiv:2309.12338). arXiv. <https://doi.org/10.48550/arXiv.2309.12338>
- Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626. <https://doi.org/10.1007/s11229-008-9435-2>
- Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive

models. *Decision Support Systems*, 51(1), 141–154.

<https://doi.org/10.1016/j.dss.2010.12.003>

Imrie, F., Davis, R., & van der Schaar, M. (2023). Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. *Nature Machine Intelligence*, 5(8), 824–829. <https://doi.org/10.1038/s42256-023-00698-2>

Ivakhnenko, A. G. (1971). Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(4), 364–378. *IEEE Transactions on Systems, Man, and Cybernetics*. <https://doi.org/10.1109/TSMC.1971.4308320>

Ivanova, A. A., Hewitt, J., & Zaslavsky, N. (2021). *Probing artificial neural networks: Insights from neuroscience* (arXiv:2104.08197). arXiv.

<https://doi.org/10.48550/arXiv.2104.08197>

Janlert, L.-E., & Stolterman, E. (2017). The Meaning of Interactivity—Some Proposals for Definitions and Measures. *Human–Computer Interaction*, 32(3), 103–138.

<https://doi.org/10.1080/07370024.2016.1226139>

Janssen, R. J., Mourão-Miranda, J., & Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, 3(9), 798–808.

<https://doi.org/10.1016/j.bpsc.2018.04.004>

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold.

Nature, 596(7873), Article 7873. <https://doi.org/10.1038/s41586-021-03819-2>

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. <https://doi.org/10.48550/arXiv.2001.08361>

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2022). Towards Causal Algorithmic Recourse. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 139–166). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_8

Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, 68, 68–79.

<https://doi.org/10.1016/j.shpsa.2018.01.011>

Kästner, L., & Haueis, P. (2021). Discovering Patterns: On the Norms of Mechanistic Inquiry. *Erkenntnis*, 86(6), 1635–1660. <https://doi.org/10.1007/s10670-019-00174-7>

Kästner, L., Langer, M., Lazar, V., Schomacker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 169–175.

<https://doi.org/10.1109/REW53955.2021.00031>

Keil, F. (2006). Explanation and Understanding. *Annual Review of Psychology*, 57(1), 227–254. <https://doi.org/10.1146/annurev.psych.57.102904.190100>

- Keil, F. (2019). *How Do Partial Understandings Work?* (pp. 191–208). <https://doi.org/10.1093/oso/9780190860974.003.0010>
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>
- Kelp, C. (2015). Understanding phenomena. *Synthese*, 192(12), 3799–3816.
- Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, 15(1), 5–12. <https://doi.org/10.1002/wps.20292>
- Kendler, K. S., & Gyngell, C. (2020). Multilevel interactions and the dappled causal world of psychiatric disorders. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 25–45). Oxford University Press. <https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697>
- Keogh, E., & Mueen, A. (2017). Curse of Dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_192
- Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press. <https://doi.org/10.1017/9781108164276>
- Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (Un)reliability of Saliency Methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 267–280). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_14
- Knuth, D. E. (1974). Structured Programming with go to Statements. *ACM Computing Surveys*, 6(4), 261–301. <https://doi.org/10.1145/356635.356640>
- Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368. <https://doi.org/10.1109/RE.2019.00046>
- Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., ... Meisenzahl, E. (2021). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry*, 78(2), 1–15. <https://doi.org/10.1001/jamapsychiatry.2020.3604>
- Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology*, 33(3), 487–502. <https://doi.org/10.1007/s13347-019-00372-9>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.

https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511498909>

Lange, M. (2017). *Because without cause: Non-causal explanations in science and mathematics*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780190269487.001.0001>

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sasing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473.
<https://doi.org/10.1016/j.artint.2021.103473>

Lawler, I., & Sullivan, E. (2021). Model Explanation Versus Model-Induced Explanation. *Foundations of Science*, 26(4), 1049–1074. <https://doi.org/10.1007/s10699-020-09649-1>

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. <https://doi.org/10.1038/nature14539>

Lederer, J. (2021). Activation Functions in Artificial Neural Networks: A Systematic Overview. *ArXiv*.

Legg, S., & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*, 17–24.

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10), 2017–2031.
https://doi.org/10.1162/jocn_a_01544

Lindsay, G. W., & Bau, D. (2023). Testing methods of neural systems understanding. *Cognitive Systems Research*, 82, 101156. <https://doi.org/10.1016/j.cogsys.2023.101156>

Lipton, P. (2008). Understanding Without Explanation. In H. W. D. Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific Understanding: Philosophical Perspectives* (pp. 43–63). University of Pittsburgh Press.

Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.
<https://doi.org/10.1145/3236386.3241340>

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464–470. <https://doi.org/10.1016/j.tics.2006.08.004>

Lombrozo, T. (2016). Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences*, 20(10), 748–759. <https://doi.org/10.1016/j.tics.2016.08.001>

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99(2), 167–204. <https://doi.org/10.1016/j.cognition.2004.12.009>

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research

directions. *Information Fusion*, 106, 102301.

<https://doi.org/10.1016/j.inffus.2024.102301>

López-Rubio, E., & Ratti, E. (2021). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, 198(4), 3131–3156.

<https://doi.org/10.1007/s11229-019-02271-0>

Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023). Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy. *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 337–342.

<https://doi.org/10.1109/REW57809.2023.00063>

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, 13(3), e1493. <https://doi.org/10.1002/widm.1493>

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT.

Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, 170(1), 169–190. <https://doi.org/10.1007/s11229-008-9366-y>

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in Translation: Contextualized Word Vectors. *Advances in Neural Information Processing Systems*, 30. <https://papers.nips.cc/paper/2017/hash/20c86a628232a67e7bd46f76fba7ce12-Abstract.html>

McCarthy, J. (1979). Ascribing Mental Qualities to Machines. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955. *AI Magazine*, 27(4), 12–14. <https://doi.org/10.1609/aimag.v27i4.1904>

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/BF02478259>

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

Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy*, 17(1), 55–70. <https://doi.org/10.1023/A:1012990030867>

Mitchell, T. M. (2013). *Machine learning* (Nachdr.). McGraw-Hill.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. <https://doi.org/10.1145/3287560.3287574>

Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, 3(3), 869–877. <https://doi.org/10.1007/s43681-022-00209-w>

Nagel, E. (1982). *The structure of science* (Reprinted). Routledge and Kegan Paul.

Nelson, R. J. (1978). The Competence-Performance Distinction in Mental Philosophy. *Synthese*, 39(3), 337–381.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126.
<https://doi.org/10.1145/360018.360022>

Nguyen, A., Yosinski, J., & Clune, J. (2019). Understanding Neural Networks via Feature Visualization: A Survey. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 55–76). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_4

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI Act. *AI & SOCIETY*.
<https://doi.org/10.1007/s00146-023-01723-z>

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(1), 13.
<https://doi.org/10.1007/s10676-022-09632-3>

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3), e00024.001.
<https://doi.org/10.23915/distill.00024.001>

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, 2(11), e7.
<https://doi.org/10.23915/distill.00007>

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010.
<https://doi.org/10.23915/distill.00010>

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>

Pearl, J., & Mackenzie, D. (2020). *The book of why: The new science of cause and effect* (First trade paperback edition). Basic Books.

Perolat, J., De Vylder, B., Hennes, D., Tarassov, E., Strub, F., de Boer, V., Muller, P., Connor, J. T., Burch, N., Anthony, T., McAleer, S., Elie, R., Cen, S. H., Wang, Z., Gruslys, A., Malysheva, A., Khan, M., Ozair, S., Timbers, F., ... Tuyls, K. (2022). Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623), 990–996. <https://doi.org/10.1126/science.add4679>

Piltaver, R., Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, 62, 333–346.
<https://doi.org/10.1016/j.eswa.2016.06.009>

Potochnik, A. (2016). *Scientific Explanation: Putting Communication First*. 721–732.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52.
<https://doi.org/10.1145/3411764.3445315>

Prince, S. J. D. (2023). *Understanding deep learning*. The MIT Press.

Pu, X., & Kay, M. (2018). The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Paper. *2018 IEEE Evaluation and Beyond*

- *Methodological Approaches for Visualization (BELIV)*, 37–45.

<https://doi.org/10.1109/BELIV.2018.8634103>

Răz, T. (2023). Methods for identifying emergent concepts in deep neural networks. *Patterns*, 4(6), 100761. <https://doi.org/10.1016/j.patter.2023.100761>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Article 1. <https://doi.org/10.1609/aaai.v32i1.11491>

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>

Roberts, D. A., & Yaida, S. (2022). *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press. <https://doi.org/10.1017/9781009023405>

Rosenberg, N. (1982). *Inside the Black Box: Technology and Economics* (SSRN Scholarly Paper 1496197). <https://papers.ssrn.com/abstract=1496197>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (pp. 318–362). MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Parallel distributed processing models of schemata and sequential thought processes. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2, pp. 216–271). MIT Press.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. <http://aima.cs.berkeley.edu/>

Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: Toward clinical applications. *Current Opinion in Neurobiology*, 55, 152–159. <https://doi.org/10.1016/j.conb.2019.02.006>

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Salmon, W. C., Jeffrey, R. C., & Greeno, J. G. (1971). *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press. <https://doi.org/10.2307/j.ctt6wrd9p>

- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229. IBM Journal of Research and Development. <https://doi.org/10.1147/rd.33.0210>
- Sarkar, A., Vijaykeerthy, D., Sarkar, A., & Balasubramanian, V. N. (2022). A Framework for Learning Ante-hoc Explainable Models via Concepts. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10276–10285. <https://doi.org/10.1109/CVPR52688.2022.01004>
- Schleidgen, S., Friedrich, O., Gerlek, S., Assadi, G., & Seifert, J. (2023). The concept of “interaction” in debates on human–machine interaction. *Humanities and Social Sciences Communications*, 10(1), 551. <https://doi.org/10.1057/s41599-023-02060-8>
- Schmidhuber, J. (2022). *Annotated History of Modern AI and Deep Learning* (arXiv:2212.11279). arXiv. <https://doi.org/10.48550/arXiv.2212.11279>
- Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3613904.3642621>
- Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3), 413–423. <https://doi.org/10.1016/j.neuron.2020.07.040>
- Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple Meanings, Goals, and Methods. *Proceedings of the 2023 AAIL/ACM Conference on AI, Ethics, and Society*, 715–722. <https://doi.org/10.1145/3600211.3604693>
- Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7(1), 1–10. <https://doi.org/10.1057/s41599-020-00614-8>
- Setiya, K. (2003). Explaining Action. *The Philosophical Review*, 112(3), 339–393.
- Sharma, A., & Verbeke, W. J. M. I. (2020). Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). *Frontiers in Big Data*, 3. <https://www.frontiersin.org/articles/10.3389/fdata.2020.00015>
- Sheu, Y.-H. (2020). Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry*, 11, 551299. <https://doi.org/10.3389/fpsy.2020.551299>
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589–1604. <https://doi.org/10.1109/JBHI.2017.2767063>
- Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. Interface agents. *Interactions*, 4(6), 42–61. <https://doi.org/10.1145/267505.267514>
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359. <https://doi.org/10.1038/nature24270>

- Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. <https://doi.org/10.1145/3531146.3534639>
- Srećković, S., Berber, A., & Filipović, N. (2022). The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. *Minds and Machines*, 32(1), 159–183. <https://doi.org/10.1007/s11023-021-09575-6>
- Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, 44(3), 510–515. <https://doi.org/10.1016/j.shpsa.2012.12.005>
- Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1), 109–133. <https://doi.org/10.1093/bjps/axz035>
- Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (Second edition). The MIT Press.
- Swanepoel, D. (2021). Does Artificial Intelligence Have Agency? In R. W. Clowes, K. Gärtner, & I. Hipólito (Eds.), *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts* (pp. 83–104). Springer International Publishing. https://doi.org/10.1007/978-3-030-72644-7_4
- Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems* (arXiv:1806.07552). arXiv. <https://doi.org/10.48550/arXiv.1806.07552>
- Trültzsch-Wijnen, C. W. (2020). Theoretical Approach to the Concepts of Competence and Performance. In C. W. Trültzsch-Wijnen (Ed.), *Media Literacy and the Effect of Socialization* (pp. 93–126). Springer International Publishing. https://doi.org/10.1007/978-3-030-56360-8_4
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(October), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, 59(4), S251–S278.
- Van Camp, W. (2014). Explaining understanding (or understanding explanation). *European Journal for Philosophy of Science*, 4(1), 95–114. <https://doi.org/10.1007/s13194-013-0077-y>
- Van Fraassen, B. C. (1980). *The scientific image*. Clarendon press.
- Vasilyeva, N., Wilkenfeld, D., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review*, 24(5), 1436–1450. <https://doi.org/10.3758/s13423-017-1275-y>
- Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- Vredenburg, K. (2022). The Right to Explanation*. *Journal of Political Philosophy*, 30(2), 209–229. <https://doi.org/10.1111/jopp.12262>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3063289>

- Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, 14(3), 270–277. <https://doi.org/10.1002/wps.20238>
- Wampold, B. E., Imel, Z. E., Bhati, K. S., & Johnson-Jennings, M. D. (2007). Insight as a Common Factor. In *Insight in psychotherapy* (pp. 119–139). American Psychological Association. <https://doi.org/10.1037/11532-006>
- Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37. <https://doi.org/10.2478/jagi-2019-0002>
- Wermter, S., & Sun, R. (2000). An Overview of Hybrid Neural Systems. In S. Wermter & R. Sun (Eds.), *Hybrid Neural Systems* (pp. 1–13). Springer. https://doi.org/10.1007/10719871_1
- Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, 190(6), 997–1016. <https://doi.org/10.1007/s11229-011-0055-x>
- Wilkenfeld, D. A. (2014). Functional explaining: A new approach to the philosophy of explanation. *Synthese*, 191(14), 3367–3391. <https://doi.org/10.1007/s11229-014-0452-z>
- Wilkenfeld, D. A. (2017). MUDdy understanding. *Synthese*, 194(4), 1273–1293. <https://doi.org/10.1007/s11229-015-0992-x>
- Wilkenfeld, D. A., & Hellmann, J. K. (2014). Understanding beyond grasping propositions: A discussion of chess and fish. *Studies in History and Philosophy of Science*, 48, 46–51. <https://doi.org/10.1016/j.shpsa.2014.09.003>
- Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., Cadena, S. A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., Sinz, F. H., & Tolias, A. S. (2023). *Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization* (p. 2023.05.12.540591). bioRxiv. <https://doi.org/10.1101/2023.05.12.540591>
- Wimsatt, W. C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy Supplementary Volume*, 20, 207–274. <https://doi.org/10.1080/00455091.1994.10717400>
- Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial Intelligence*, 170(18), 1256–1258. <https://doi.org/10.1016/j.artint.2006.10.011>
- Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365. <https://doi.org/10.1038/nn.4244>
- Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1), 3770. <https://doi.org/10.1038/s41467-019-11786-6>
- Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288. <https://doi.org/10.1007/s13347-019-00382-7>

Chapter 1: Introduction

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 107–115. <https://doi.org/10.1145/3446776>

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5), Article 5. <https://doi.org/10.3390/electronics10050593>

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), e2014196118. <https://doi.org/10.1073/pnas.2014196118>

Chapter 2:

Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-based Approach

This chapter was published as “Speith, T., Crook, B., Mann, S., Schomäcker, A., & Langer, M. (2024). Conceptualizing understanding in explainable artificial intelligence (XAI): an abilities-based approach. *Ethics and Information Technology*, 26(2), 40.” Slight changes have been made to fit the dissertation format.

2.0 Abstract

A central goal of research in explainable artificial intelligence (XAI) is to facilitate human understanding. However, understanding is an elusive concept that is difficult to target. In this paper, we argue that a useful way to conceptualise understanding within the realm of XAI is via certain human abilities. We present four criteria for a useful conceptualisation of understanding in XAI and show that these are fulfilled by an abilities-based approach: First, thinking about understanding in terms of specific abilities is motivated by research from numerous disciplines involved in XAI. Second, an abilities-based approach is highly versatile and can capture different forms of understanding important in XAI application contexts. Third, abilities can be operationalised for empirical studies. Fourth, abilities can be used to clarify the link between explainability, understanding, and societal desiderata concerning AI, like fairness and trustworthiness. Conceptualising understanding as abilities can therefore support interdisciplinary collaboration among XAI researchers, provide practical benefit across diverse XAI application contexts, facilitate the development and evaluation of explainability approaches, and contribute to satisfying the societal desiderata of different stakeholders concerning AI systems.

2.1 Introduction

Many artificial intelligence (AI) systems remain opaque to the stakeholders who interact with them (Burrell, 2016; Mann et al., 2023). This hinders the fulfilment of societal desiderata such as trust, fairness, and accountability (Langer et al., 2021c). For example, when companies use AI systems for hiring decisions, it can be problematic if those systems remain a black box to the human resource managers who use them or the applicants who are subject to their decisions. In light of these concerns, it is increasingly recognised that an ethically acceptable integration of AI into society will require humans to *understand* AI systems. To address this issue, the growing, interdisciplinary research field of explainable AI (XAI) tries to render AI systems understandable (Barredo Arrieta et al., 2020; Langer et al., 2021c; Páez, 2019).⁵⁵

However, a roadblock to progress in XAI is a lack of clarity about what it means to understand an AI system. This creates several problems. First, as an interdisciplinary endeavour (Langer et al., 2021a, 2021c; Páez, 2019; Miller et al., 2017; Lipton, 2018), XAI requires collaboration across disciplines. When ethicists and legal scholars demand that we need to understand AI systems to ensure fairness or legal liability (e.g., Vredenburg, 2022; Deeks, 2019) or when computer scientists present a new method to enable better understanding of AI systems (e.g., Ribeiro et al., 2018; Lapuschkin et al., 2016), these parties need to agree on what they mean when they refer to “understanding” AI. Second, appeals to “sufficient understanding” of AI-based systems (as in Article 14 of the proposal for a European AI Act)⁵⁶ do not reflect how requirements for understanding vary depending on the specific application context (e.g., on the *stakeholder* who aims to understand AI, the *system-related aspect* they want to understand, and the *desideratum* which motivates their need for understanding).⁵⁷ For example, in the case of an AI hiring system, an applicant may want to understand why their application was rejected

⁵⁵ Note that we interpret XAI broadly, as the interdisciplinary research field addressing epistemic challenges arising from the opacity of AI systems. This contrasts with a narrow construal of XAI as the development of post-hoc explanation methods (see, e.g., Rudin, 2019).

⁵⁶ <https://artificialintelligenceact.eu/the-act/>

⁵⁷ In the XAI debate, there are many possibilities for the exact aspect that is supposed to be understood (for an overview, see Chazette et al., 2021). For our argument, however, it is only important that the understanding is concerned with some aspect of a system, that is, a *system-related aspect*.

(Wachter et al., 2017), while the developer may need to understand whether the system is fair (Hutchinson & Mitchell, 2019). Third, understanding is often construed as a cognitive concept that can be analysed in terms of an agent's internal states (see, e.g., Wilkenfeld, 2013). However, accounts that are not closely associated with empirically testable measures remain an elusive target for the design and evaluation of explainability approaches. Fourth, although there is widespread agreement that a greater understanding of a system will be beneficial (Barredo Arrieta et al., 2020; Chazette et al., 2021; Langer et al., 2021a, 2021c; Speith, 2022a; Hoffman et al., 2018), it remains unclear how, exactly, the concept of understanding relates to the satisfaction of societal desiderata.

In this paper, we propose that advancing research on XAI requires a conceptualisation of understanding that is tailored to the specific needs of the field. To this end, we argue for explicating different ways of understanding⁵⁸ an AI system in terms of specific understanding-related *abilities* (e.g., assessing, predicting) a person may have with regard to that system.⁵⁹ We believe that an abilities-based conceptualisation of understanding has the potential to overcome the aforementioned problems.

The structure of this paper is as follows: Drawing on the specific needs of XAI outlined above, we begin by establishing four criteria that a useful conceptualisation of understanding in XAI should fulfil: It must be (i) motivated by prior research across relevant disciplines, (ii) versatile, (iii) operationalisable, and (iv) establish a link to the satisfaction of relevant societal desiderata. The rest of the paper is devoted to showing that abilities fulfil these criteria to a high degree. First, we demonstrate that abilities unite interdisciplinary research by documenting their importance for accounts of understanding in different disciplines. Then, we argue that abilities are versatile by presenting six clusters of abilities that capture different ways of understanding required for diverse XAI application contexts. Subsequently, we show that our proposed ability clusters are operationalisable, as reflected by prior research using abilities to evaluate

⁵⁸ We speak of *ways* or *forms* of understanding in order to remain neutral regarding debates on different types and degrees of understanding. For discussion on these issues, see, e.g., Baumberger (2014) and Baumberger et al. (2017).

⁵⁹ Unless further qualified, we will use “abilities” in the following to mean specific abilities related to understanding.

XAI methods. Next, we show how abilities can clarify the relationship between understanding and the satisfaction of societal desiderata. Finally, we address potential limitations of an abilities-based approach to understanding and conclude.

2.2 Criteria for a Useful Conceptualisation of Understanding in XAI

In philosophy, there is an extensive debate about what exactly constitutes understanding (Baumberger et al., 2017; Hannon, 2021). Similarly, different models of what it is to understand have been proposed across various disciplines (Pearl & Mackenzie, 2018; Endsley, 1995; Thórisson et al., 2016; Bloom et al., 1956). What is lacking, however, is a shared notion of understanding that is tailored to XAI. In this paper, we provide a conceptualisation of understanding that is useful specifically for the debate on explainability, and that allows for fruitful interdisciplinary research advancing the field.⁶⁰ Based on our knowledge of XAI, we specify four criteria below which a conceptualisation of understanding should fulfil in order to serve that purpose.⁶¹

Criterion 1: Interdisciplinary motivation

First, a useful conceptualisation of understanding in XAI should be *motivated across disciplines*. The research field of XAI is highly interdisciplinary, stretching from computer science via psychology and law to philosophy and the social sciences. Such interdisciplinary endeavours come at the risk of producing misunderstandings as different disciplines use concepts according to their particular needs. However, the general development of research on XAI as well as the development and assessment of individual approaches requires a smooth collaboration between the fields (Langer et al., 2021a, 2021c; Speith, 2022b). Therefore, it is crucial to align the use of important concepts.

⁶⁰ By *conceptualisation* we mean a specific account of a concept that may not capture every aspect of it but serves a particular research purpose.

⁶¹ Our criteria bear resemblance to those used in conceptual engineering research (v. Carnap, 1962; Brun, 2016). In conceptual engineering, a new term is introduced to serve a specific theoretical purpose. Well-known criteria for assessing the new term are similarity to the previous term (our Criterion 1) and fruitfulness for further research (our Criteria 2–4).

In light of the interdisciplinary nature of XAI, insisting on a conceptualisation of understanding that is particular to any one discipline would be misguided (Miller et al., 2017; Miller, 2019). Instead, it is preferable to conceptualise understanding in a way that draws from multiple disciplines and, consequently, is plausible to most researchers in the XAI debate. We thus aim to find a conceptualisation that unites research on the concept of understanding from different disciplines relevant for XAI.

Criterion 2: Versatility

The next property of a useful conceptualisation is versatility. We are looking for a conceptualisation that is sensitive to the variety of understanding-related requirements of concrete XAI application contexts. We thus need a conceptualisation of understanding that allows for differentiating the ways in which understanding can manifest. To illustrate, we refer to the following example throughout the paper:

Hiring Case Example

An AI system is used to support hiring decisions (Hickman et al., 2022; Langer et al., 2021b; Baum et al., 2022). As input, it takes application data such as résumés or results from cognitive ability tests. It then outputs a ranking of applicants. A hiring manager plays the role of a human-in-the-loop and decides, with the help of this ranking, who should be invited to interview for a position. Explainability is relevant for such a system, for instance, to ensure the system's outputs do not discriminate unfairly against any protected groups (e.g., women, people of colour, LGBTQIA*; Carvalho et al., 2019; Langer & König, 2023).

Based on this example, we can think of several XAI application contexts in which different forms of understanding might be important: The applicants might want to understand why their application was rejected (i.e., they need to understand a single output of the program). Alternatively, if we want to hold the hiring manager accountable for evaluating applicants based on the system's output, the manager might need to understand how the system uses information about applicants to determine hirability (Baum et al., 2022). Finally, a system developer, who wants to improve the system, needs to understand how the system's decision making is implemented in the code.

Clearly, then, various forms of understanding are required in different XAI application contexts. Therefore, it is important to rely on a conceptualisation of understanding that is able to accommodate the distinctions between these forms. Making such distinctions allows for greater precision and specificity when thinking about what is needed in each case. This should help researchers and practitioners spell out concrete requirements for the evaluation and design of explainability approaches by focusing attention on the form of understanding relevant for the specific application context.

Criterion 3: Operationalisability

Third, for the field of XAI to advance, it needs to adopt a conceptualisation of understanding that can be empirically evaluated. An important objective within XAI research is to assess and compare the success of different explainability approaches at inducing understanding in different contexts. To that end, the effects of an approach on different stakeholders with varying background knowledge and in diverse contexts should be measured in empirical studies. This requires a conceptualisation of understanding that allows for measurement of whether an explainability approach facilitates the required understanding of an AI system. Not all conceptualisations of understanding immediately offer themselves to such an operationalisation. For example, some philosophical accounts of understanding focus on purely cognitive aspects of understanding like the manipulation of mental representations (e.g., Wilkenfeld, 2013).⁶² For pragmatic reasons, a conceptualisation of understanding should be relatable to observable human behaviour.

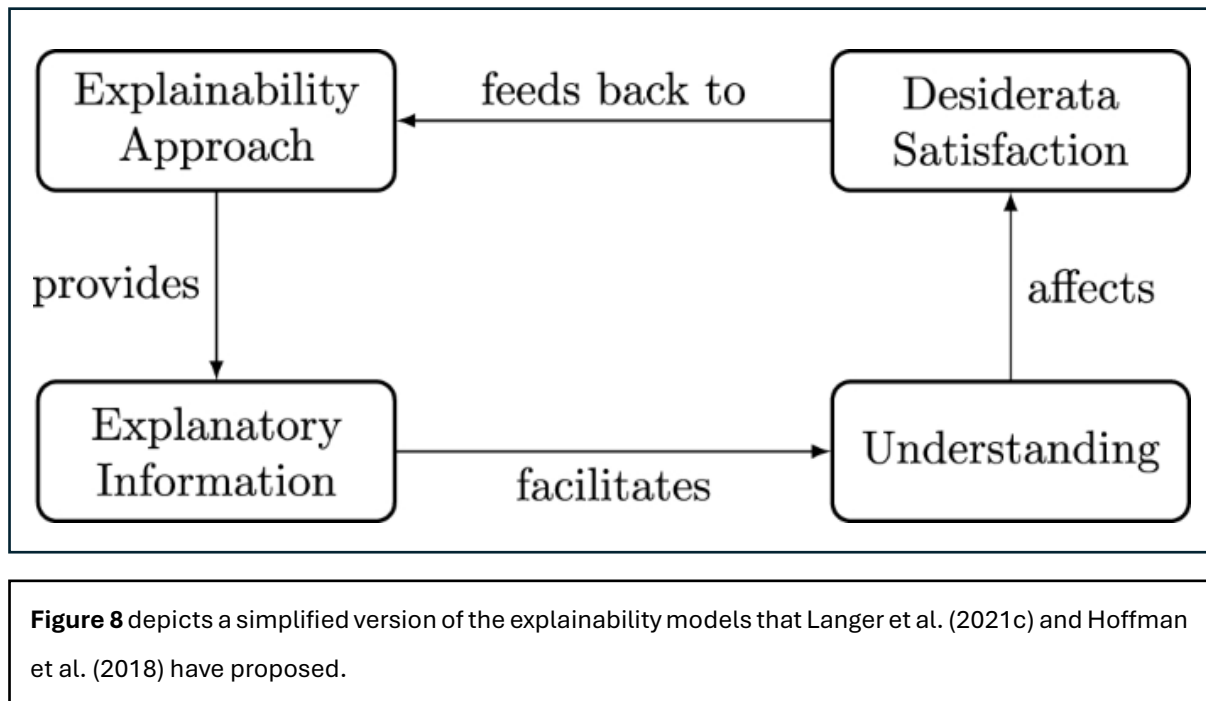
Criterion 4: Link to desiderata

The last property that we consider important for a useful conceptualisation of understanding in XAI is that it can be used to specify the relationship between explainability, understanding, and desiderata. Understanding is usually not a goal in itself for XAI, but we rather want to understand AI to ensure, e.g., fairness, accountability, or human oversight (Krishnan, 2020; Barredo Arrieta et al., 2020; Chazette et al., 2021; Langer et al., 2021a, 2021c; Speith, 2022a, 2022b; Hoffman et al., 2018). In particular,

⁶² As we are not looking to give a definition of understanding, we do not want to exclude that these accounts point to some important aspects of the nature of understanding in general.

Hoffman et al. (2018) and Langer et al. (2021c) propose similar models that outline how key concepts in XAI relate to one another (see Fig. 1). According to their models, explainability approaches provide explanatory information with the aim of facilitating people's understanding. This understanding, in turn, affects the satisfaction of desiderata. Depending on the final status of desiderata satisfaction, the explainability approach may need to be adapted or a new one chosen.

FIGURE 8: A SIMPLIFIED MODEL OF EXPLAINABILITY



Langer et al. (2021c) specify their model further by introducing the distinction between *epistemic* and *substantial* desiderata satisfaction. For substantial desiderata satisfaction, an AI system, or an element of the sociotechnical system in which it is embedded, has to have a certain property (e.g., the AI's outputs have to fulfil some mathematical fairness measure, or its user needs to be in a position to bear responsibility for its failure). For epistemic desiderata satisfaction, stakeholders require epistemic access to whether or not a desideratum is satisfied substantially (e.g., they have to know whether a system is fair). Understanding primarily leads to epistemic desiderata satisfaction. However, understanding can also support substantial desiderata satisfaction (e.g., when understanding the system supports debugging to make it fairer). Sometimes, understanding is even required for substantial desiderata satisfaction (e.g., human oversight of a system can only be ensured when certain people understand it). In

many cases, however, the exact relationship between explainability, understanding, and (epistemic or substantial) desiderata satisfaction is not clear (see, e.g., Kästner et al., 2021; Deck et al., 2024). Thus, a useful conceptualisation of understanding should serve to clarify how, exactly, desiderata can be satisfied by explainability.

In the following sections, we will argue that conceptualising understanding via abilities fulfils all of these criteria to a high degree.

2.3 Abilities Are Motivated by Research Across Disciplines

Our first criterion requires that a useful conceptualisation of understanding in XAI is both motivated by and fruitful for disciplines relevant to XAI. We demonstrate interdisciplinary plausibility of focusing on abilities by documenting convergence across various fields in utilising abilities to analyse and operationalise understanding.

In *philosophy*, one of the central aspects of understanding is considered to be the *grasping* of relationships, e.g., between the elements of an explanation, a theory, or a scientific model (Kvanvig, 2009; Baumberger, 2014; Riggs, 2003; Grimm, 2011). Grasping, in turn, is usually spelled out in terms of distinct abilities someone who understands is taken to possess (Baumberger et al., 2017). For instance, Hills (2016) lists several abilities characteristic of *understanding-why* (e.g., understanding why a job applicant was rejected). These include the ability to follow an explanation, to explain in one's own words, or to draw conclusions about relevantly similar cases. Others have suggested further abilities, e.g., the ability to answer questions about counterfactual cases (Grimm, 2011), to make predictions (de Regt, 2015), to qualitatively solve problems (Newman, 2017), to construct (scientific) models (de Regt, 2015), or to evaluate competing explanations (Khalifa, 2013).

Abilities are also central to how understanding is envisioned and studied in *cognitive psychology*. For example, Williams and Lombrozo (2010) treated the ability to reason and generalise about how object properties relate to object categories as a measure of understanding (see also Williams et al., 2010). Relatedly, Rozenblit and Keil (2002) used the ability to produce diagrammatic explanations of everyday objects as a measure of understanding, showing that people were systematically overconfident in their self-

assessed understanding. Finally, research in cognitive psychology suggests that our intuitive judgments of explanation and understanding may be ability-centered (Vasilyeva et al., 2015, 2017; Lombrozo & Carey, 2006).

In *educational psychology*, assessing students' abilities is foundational to measuring their understanding of educational content. The locus classicus of this line of research is *Bloom's Taxonomy* (Bloom et al., 1956), an influential attempt to systematise the wide variety of abilities that capture student understanding. Krathwohl (2002)'s revision of the taxonomy consists of six hierarchical levels with associated abilities: remembering, understanding, applying, analysing, evaluating, and creating.

In research on *AI and causal inference*, Pearl and Mackenzie (2018)'s Ladder of Causation relates understanding to possessing certain abilities. The ladder has three rungs: association, intervention, and counterfactuals. In another contribution from the field of AI, Thórisson et al. (2016) provide a pragmatic way of testing the understanding of any potentially intelligent agent. In this case, the authors associate understanding with an agent's possession of the following abilities (with respect to a target phenomenon): predicting, achieving goals, explaining, and re-creating.

Abilities are also used to operationalise understanding in the area of *human factors* research. For example, situation awareness, which is understanding of one's environment, concerns "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1995, p. 36). Situation awareness has three hierarchical levels; in ascending order: perception, comprehension, and projection.

Overall, numerous disciplines consider abilities to be a means of capturing understanding. Further, there is a significant overlap between the abilities referred to by different disciplines. For example, abilities such as generalising, explaining, and counterfactual reasoning are part of many conceptualisations of understanding. Additionally, many disciplines propose some kind of hierarchy of abilities, corresponding to increased degrees of understanding, according with the intuition that certain abilities are more demanding than others (Bloom et al., 1956; Pearl & Mackenzie, 2018). This convergence suggests that an abilities-based account of understanding is well-suited to serve as a unified conceptualisation for interdisciplinary research in the field of XAI.

Despite these similarities, however, the various disciplines we have discussed present different accounts of which abilities are relevant to understanding. We take this to indicate that which abilities best reflect understanding can vary depending on the context. We suggest, therefore, that XAI application contexts require a tailored account of understanding-related abilities. To this end, we next present a systematisation of abilities suited to capturing the different forms of understanding relevant to XAI application contexts.

2.4. Abilities Are Versatile

Our second criterion for a useful conceptualisation of understanding in XAI demands that it should be versatile enough to capture the variety of forms of understanding that may be required in XAI application contexts. To this end, in this section, we integrate and systematise the insights gleaned from our interdisciplinary review and propose six clusters of abilities (see Fig. 9) that reflect differing ways of understanding: *recognising*, *assessing*, *predicting*, *intervening*, *explaining*, and *designing*.⁶³ We intend these six clusters to reflect existing work and, crucially, to capture the ways of understanding that are relevant across XAI application contexts.

Recognising

A foundational ability cluster is *recognising*. Recognising is apprehending what is signified by a piece of information such as a text, an image, or a number. As such, in many cases, recognising is an undemanding ability possessed, e.g., by any person reading a text consisting of familiar words in their own language. However, while recognising may seem trivial, many representational formats common to the AI domain, such as vector representations, correlation matrices, and saliency maps, may be unrecognisable to laypeople, especially if they are presented without adequate labelling or accompanying description (Langer et al., 2021c; Franconeri et al., 2021; Speith, 2022a). In general,

⁶³ We call them “clusters” to reflect that they represent umbrella concepts whose concrete operationalization may change depending on the context. For the same reason, we entreat readers to focus on the substantive descriptions of the abilities, rather than their labels.

recognising is crucial for basic use of AI systems and often serves as an important prerequisite for achieving more demanding forms of understanding.

FIGURE 9: A LOOSE HIERARCHY OF CLUSTERS OF ABILITIES

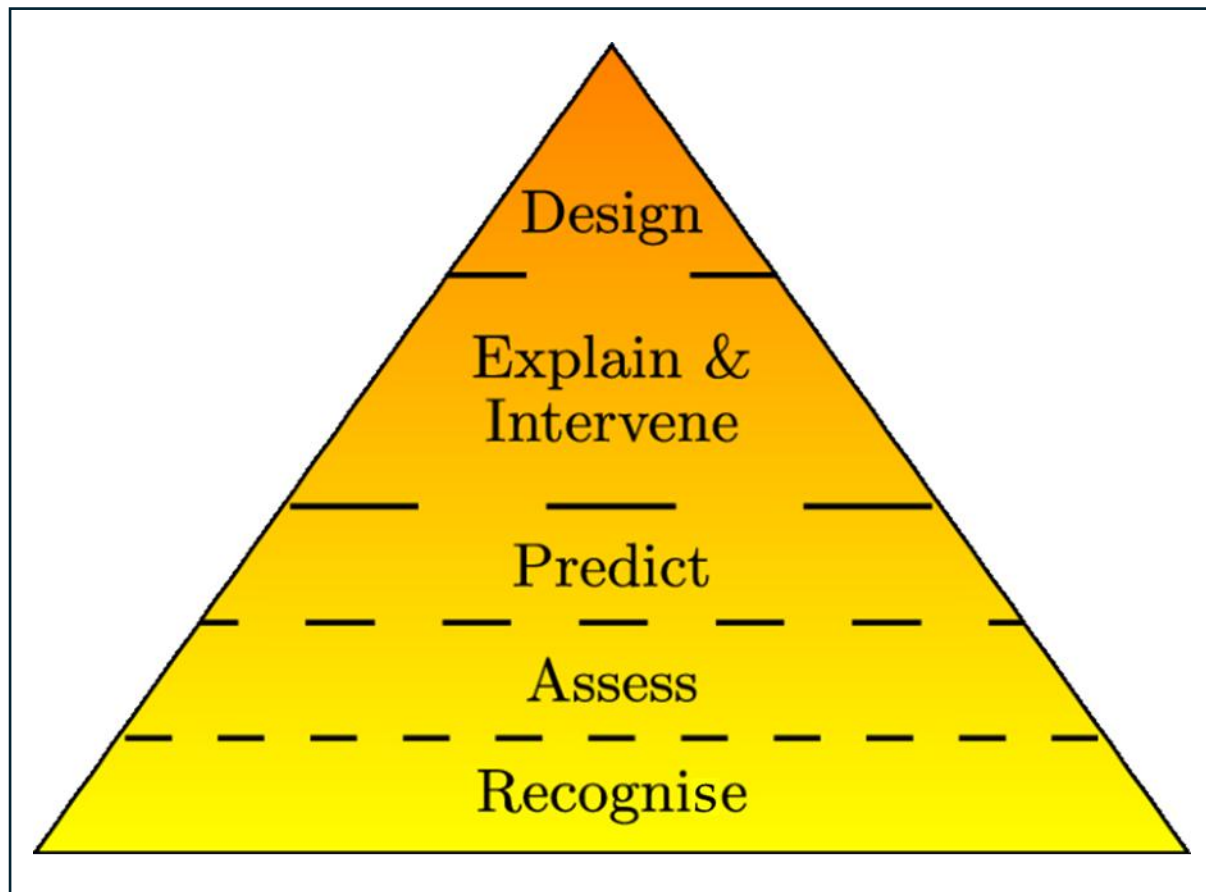


Figure 9 depicts our proposed clusters of abilities relevant to characterizing different ways of understanding in XAI. The hierarchical structure reflects that some abilities tend to be more demanding than others. However, how demanding the acquisition of an ability is in a specific case is also influenced by contextual factors.

For example, consider the hiring manager in our application scenario. To use the system in the first place, the manager needs to recognise the system's outputs as ascribing a certain rank to an applicant (e.g., that a woman named April is ranked 10th). This limited understanding is sufficient for a person to follow the recommendations of the decision support system, even if nothing else about the system is known. In this example, the output being presented in an accessible way would suffice to induce this ability. Insight into the inner workings of the system is not required.

Abilities similar to what we call recognising can be found in many disciplines, always on a comparably low hierarchical level. For instance, perceiving in situational awareness (Endsley, 1995) and remembering in Bloom’s Taxonomy (Krathwohl, 2002) are close to what we mean by recognising.

Assessing

The next ability cluster we specify is *assessing*. Once one recognises what a system-related aspect (e.g., an output) refers to, the next step is an (initial) assessment of the adequacy of that aspect. This ability is characterised, then, by accurately judging when a system-related aspect is inadequate, unreasonable, or has an unacceptable risk of being so.

In some cases, recognising an output, in combination with basic background knowledge, will suffice for assessing.⁶⁴ For instance, an AI system that classifies a human being as a gorilla is obviously flawed (Garcia, 2016). In other cases, AI systems behave inadequately in less straightforward ways, particularly when they are intended to satisfy complex and multifaceted desiderata such as fairness. In such cases, information about the reasoning processes of the AI may be needed to assess the adequacy of an output or some other system-related aspect (Baum et al., 2022).

Assessing, as we construe it, relates to the ability to evaluate that occupies the fifth level in the revision of Bloom’s Taxonomy (Krathwohl, 2002) and constitutes one facet of comprehension in situation awareness (Endsley, 1995). Being able to assess a system’s output is a prerequisite for taking appropriate action in response. This could mean demanding further explanation, deciding not to use a system, or bringing a legal challenge against a system’s operators. Assessing is typically more demanding than recognising, as one can recognise what a system’s output denotes without being able to judge whether it is appropriate or satisfactory.

⁶⁴ By “background knowledge” we mean any (propositional or procedural) knowledge a person possesses prior to their interaction with an AI system. Background knowledge plays a crucial role, as the explanations provided by explainability approaches always need to match the stakeholder’s previous background knowledge in order to support the abilities we discuss.

Predicting

The third ability cluster we propose is *predicting*. We characterise predicting as the anticipation of a system's behaviour. As such, this ability is exemplified by accurately estimating what a system will output given some input. For example, will a job applicant be ranked favourably or not? How will a particular image be classified? A person possesses this ability with respect to a system if they are able to reliably predict (roughly) what it will do.

Consider our hiring scenario. Assume that April contacts a consultant to find jobs for her. Drawing on their experience, the consultant can predict that April will be ranked favourably at a company employing the hiring system. Thus, they encourage her to apply there. Explainability approaches that highlight the relevance of particular features may be effective at supporting this ability of the consultant (e.g., LIME, see Ribeiro et al., 2016 or Anchors, see Ribeiro et al., 2018; see also Speith, 2022b).

Predicting is a broad and graded ability that is related to multiple notions from our interdisciplinary survey. For example, it corresponds closely to what Pearl and Mackenzie (2018) call association. In line with Pearl and Mackenzie (2018)'s Ladder of Causation, this ability does not necessarily require causal knowledge about how the input is transformed into the output (as also noted by Knüsel & Baumberger, 2020). In practice however, accurate prediction often requires augmenting observations of input-output pairs with insights into how the transformation takes place.

Further notions, such as projection in human factors or generalising in cognitive psychology, are also kinds of prediction (see, e.g., Williams & Lombrozo, 2010). For example, generalising is predicting the properties of an object, given knowledge of its class. What unifies predicting, despite its heterogeneity, is the mental simulation of a mapping from inputs to outputs. Predicting is distinct from assessing. One can assess a system's behaviour, that is, notice when it produces inadequate outputs, without being able to predict what those outputs will be. In this sense predicting is more intimately related to knowledge of and experience with a particular system than assessing.

Intervening

The next cluster of abilities we identify is *intervening*. Intervening is acting on a system so as to accomplish a goal. It therefore requires finding the means to achieve some specified ends. Intervention can have different objects. On the one hand, intervention can be directed at the *information* a system processes, e.g., when tailoring input features to achieve a certain result. On the other hand, intervention can aim at altering how the system *processes* information, e.g., to achieve some desired system property such as monotonicity.

In our hiring case, the first type of intervention could be achieved by the applicant. Aiming to improve her ranking, April might tailor her application by including specific information that is known to be influential. The second kind of intervention would largely be exclusive to developers, namely, changing the inner workings of the system in order to, e.g., improve the system's accuracy or fairness. Notice that being able to intervene in the latter way requires knowledge about the properties of the system and how it implements the transformation of inputs into outputs. As such, this ability typically depends upon significant background knowledge. However, it may also be supported by model-specific explainability approaches that shed light on how different system properties and components contribute to overall behaviours (e.g., Lapuschkin et al., 2016; Stock & Cissé, 2018; see also Speith, 2022b).

Intervening is related to de Regt (2015)'s view that scientific understanding of phenomena is characterised by being able to use a model to control its behaviour as well as Thórisson et al. (2016)'s notion of achieving goals. Intervening can be distinguished from predicting, which requires forward simulation from a set of initial conditions to their likely consequences. Intervening, in contrast, depends on reasoning backwards from a chosen outcome to the changes in the conditions that would produce it. While learning to predict is possible from pure observation of input-output behaviour without specific knowledge of how a system works at the level of components (Knüsel & Baumberger, 2020), intervention typically demands knowledge of which variables need to be changed to alter the system's behaviour.

Explaining

Our fifth cluster is *explaining*. Explaining can mean providing causal information about an occurrence or giving a description of how something works in terms of its components and their interactions (Rozenblit and Keil, 2002; Halpern & Pearl, 2005).⁶⁵ As with other abilities, explaining comes in degrees. Partial explanations describe some of the causes, components, and interactions relevant to a phenomenon, while complete explanations describe them all.

To exemplify explaining, assume that the hiring system in our example is based on a decision tree classifier. One can partially explain the system's output by pointing to influential input features (e.g., educational attainment). A complete explanation, however, would also require describing the tree's nodes, how they are organised, and how they interact to produce the system's outputs. Providing detailed explanations is a demanding ability which typically requires significant expertise with respect to the target of explanation (Rozenblit and Keil, 2002). Information provided by explainability approaches may be crucial in supporting the development of the ability to explain. For example, model-specific approaches to explainability often attempt to characterise how components contribute to the functioning of a system (e.g., Lapuschkin et al., 2016; Cammarata et al., 2020; see also Speith, 2022b).

Our notion of explaining coheres with Rozenblit and Keil (2002)'s operationalisation of understanding and is close to accounts in the philosophy of science which require, e.g., being able to specify the causal components and relationships relevant to some phenomenon (Newman, 2017; Halpern & Pearl, 2005; Strevens, 2008). Explaining closely relates to intervening, as both clusters presuppose some knowledge of inputs or processes in a system. However, while intervening does not require the ability to articulate this knowledge, explaining does. On the other hand, explaining may not require the technical know-how needed for altering inputs or changing the system. For this reason, we place these abilities on the same level in our hierarchy.

⁶⁵ This holds at a practically relevant level of abstraction. That is, explaining an artificial neural network might require specifying how input data and learned weights interact to produce outcomes, but not how those properties are implemented in machine language (Craver & Kaplan, 2020; Potochnik, 2010).

Designing

The final cluster of abilities we identify, constituting comprehensive understanding, is *designing*. Designing typically applies to whole systems and requires being able to specify which components are needed, how they are organised, and how they function to produce outcomes satisfying relevant requirements. As such, being able to design an AI system from scratch, such that it satisfies specific constraints on, say, accuracy and fairness, would constitute a deep understanding of that system and the problem it addresses.

In our hiring case, a model developer might be tasked with designing an appropriate system. This could require an iterative process of feature engineering to construct meaningful and robust features that support fair and accurate decisions (Rudin, 2019). Note that, since we take designing to require specifying *how* a system functions, training black-box models like deep neural networks does not suffice for this ability.

Designing is related to Thórisson et al. (2016)'s notion that fully understanding a phenomenon means being able to re-create it and de Regt (2015)'s view that scientific understanding involves the construction of models. Additionally, it resembles the final level of Bloom's revised taxonomy: creating (Krathwohl, 2002). This form of understanding requires being able to put elements of knowledge together to develop a novel product.

Notice the distinction between intervening and designing. The former requires altering a system in some way, perhaps by tweaking the values of its parameters, while the latter requires actually *specifying* the components of a system and how they are organised. Further, designing is related to, but more demanding than explaining. While partial explanations are often sufficient for the satisfaction of desiderata, incomplete designs are of limited value. However, a complete diagrammatic explanation of a system including all components and their interactions essentially is a design for that system (see also Cammarata et al., 2020).

Note that we depict our clusters in the form of a pyramid to reflect that some of them tend to be more demanding than others (see Fig. 9). However, how demanding it is to acquire each ability is also influenced by contextual factors. For instance, assessing whether a single prediction relied on protected attributes may be easier than assessing whether the

system satisfies fairness desiderata at a global level. Furthermore, there are scenarios where someone might possess an ability higher up in the hierarchy without possessing the abilities below it. For example, in absence of relevant domain knowledge, the ability to explain does not necessarily entail the ability to assess whether an output was appropriate. Crucially, explainability approaches should not always aim to produce the abilities we place higher up in the pyramid. Rather, they ought to facilitate whatever form of understanding is required to satisfy downstream desiderata (see Chapter 2.6).

2.5 Abilities Are Operationalisable

The next criterion of a useful conceptualisation of understanding in XAI is that it should be operationalisable. While the interdisciplinary review of understanding in Chapter 2.3 has already demonstrated that abilities can be operationalised for empirical studies (e.g., in cognitive or educational psychology), we now show that they can also be used in XAI. More specifically, we show that the abilities we describe in our clusters have already been used in studies to evaluate explainability approaches; thus, the clusters we propose are operationalisable.

Recognising

Recognising is a fundamental ability that is often taken for granted. As such, most studies presuppose this ability by choosing participants that are familiar with the representational format used, or are primed beforehand. However, Vilone and Longo (2021)'s review of explainability approaches summarises the strengths and weaknesses of different representational formats (e.g., numeric, rule-based, visual, textual), incorporating concerns about whether certain stakeholders can really recognise some representational formats (e.g., laypeople may struggle with mathematical formulae as explanations; see also Speith, 2022a, 2022b).

Assessing

Ribeiro et al. (2016, 2018) have proposed two explainability approaches, LIME and Anchors, both of which they evaluated via measuring abilities. In one of their studies evaluating LIME, participants were asked to choose between several equally accurate classifiers which relied on different classification strategies. This procedure tests for participants' ability to assess the quality of different systems. Another example can be

found with Piltaver et al. (2014), who also used an assessing task to test the comprehensibility of classification trees. In particular, they had participants validate (parts of) the trees. Furthermore, Lapuschkin et al. (2016) demonstrated the efficacy of their Layer-wise Relevance Propagation (LRP) approach by using it to support assessing whether image classifiers were detecting meaningful features or artifacts. See Stock and Cissé (2018) and Allahyari and Lavesson (2011) for further tasks that can be attributed to our assessing cluster.

Predicting

Piltaver et al. (2014)’s study also measured prediction abilities. In particular, participants in their study had to derive the output of a decision tree given an input. Similarly, Huysmans et al. (2011) evaluated understanding of various classification models by having participants predict these models’ classifications on previously unseen instances. See Alqaraawi et al. (2020), Poursabzi-Sangdeh et al. (2021), and Ribeiro et al. (2018) for further tasks from our prediction cluster.

Intervening and explaining

In the evaluation of LIME, Ribeiro et al. (2016) tasked participants to a) find a classifier’s source of failure and b) improve a classifier after receiving explanations. The former task assessed the participants’ ability to explain the failure of a system, while the latter evaluated their ability to intervene on the classifier. Piltaver et al. (2014) also conducted a study belonging to the intervening cluster. They tasked participants with identifying which attribute values in a decision tree had to be changed to alter a classification. Another explaining task can be found in Tullio et al. (2007).

Designing

Designing is usually so challenging that it is not addressed with empirical studies. Pragmatically, however, designing is the goal for model developers tasked with constructing inherently understandable models. This is exactly what Rudin (2019) calls for: She argues that in most problem domains, given sufficient time and developer expertise, such models can be designed (see also Chapter 4). While designing, to the best of our knowledge, has not been tested empirically, it is evaluated by the process of constructing inherently understandable models, e.g., when checking whether a model

satisfies certain functional requirements. Further, designing is sometimes a goal of scientific investigations. For example, Cammarata et al. (2020) attempt to reverse engineer a deep neural network so that it can be re-implemented from scratch (i.e., by setting the connection weights between neurons in the network manually). In principle, one could imagine using this task in a study.

Overall, we believe that the work cited above confirms not only that abilities can be operationalised (i.e., that they can be used to design and carry out empirical studies to measure understanding and thereby evaluate explainability approaches), but also that our proposed clusters are useful to this end.

2.6 Abilities Can Be Linked to Desiderata

Turning to the final criterion, a useful conceptualisation of understanding in XAI should illuminate the intricate relationship between explainability, understanding, and desiderata. As the models of Langer et al. (2021c) and Hoffman et al. (2018) illustrate, it is commonly assumed in XAI that explainability approaches, by facilitating understanding, are supposed to satisfy desiderata (e.g., fairness, trust, or safety; see Fig. 8), either epistemically or substantially. However, it is often unclear how, exactly, this is supposed to work (see, e.g., Kästner et al., 2021; Deck et al., 2024). We claim that by using the ability clusters we propose in Chapter 2.4, one can describe more precisely the relationship between explainability, understanding, and desiderata. In particular, abilities can be used to specify the requirements for epistemic or substantial desiderata satisfaction in a given situation.

Example 1: Fairness

To exemplify the connection between abilities and desiderata, let us first consider a desideratum prevalent in the hiring case: *fairness*. In the XAI literature, it is often assumed that explainability is important for fairness (for a review, see Deck et al., 2024); however, as of yet, there is little agreement on how, exactly, a better understanding of the system can contribute to fairness. Therefore, anyone who wants to improve fairness via XAI needs to be clear about how they expect a better understanding of the system to affect fairness. In cases like this, our account proves useful.

Consider April again, whose application was rejected. In this situation, April may want to know whether she was treated unfairly. Thus, she requires epistemic access to whether the system's decision was fair. For this kind of epistemic desideratum satisfaction, April would need to be able to *assess* what led the system to the specific output "rejection," specifically concerning the influence of protected attributes. Now consider an external auditor, Sam, whose role is to check whether the hiring support system adheres to a particular notion of fairness (e.g., subgroup parity, Hutchinson & Mitchell, 2019). This is another case that aims for epistemic desideratum satisfaction, which, however, requires different abilities. For Sam to gain epistemic access to whether the system is fair overall, he would not only need the ability to assess certain statistical properties of the system's behaviour, but also to reliably *predict* that the system's outputs, overall, will continue to adhere to the chosen notion of fairness.

Other ability clusters can specify how understanding can contribute to the substantial satisfaction of the fairness desideratum: If the system under inspection turns out *not* to be fair in the desired sense, this may prompt a requirement for further abilities. For instance, to rectify an unfair system, a developer would need the ability to *intervene* on the system. This would require changing its internal processing in a way that ensures fair behaviour in the future. One can imagine even more demanding abilities being required to ensure system fairness. For instance, Rudin (2019) rejects so-called black-box models in high-stakes scenarios because it is hard to determine whether their decision making is fair. Instead, she prefers ante-hoc explainable models (e.g., logistic regression or decision lists). If ante-hoc explainability was the only way to ensure system fairness, this would require a complete description or explanation of the system's decision-making processes, as well as demonstration that those processes adhere to the chosen notion of fairness. As noted in Chapter 2.4, the ability to give such a complete explanation is tantamount to the ability to *design* the system.

The presented examples demonstrate that the relationship between our ability clusters and desiderata is so close that, in some cases, specifying which ability a stakeholder (or several stakeholders) needs to gain also describes the requirements for desiderata satisfaction. For both April and Sam, gaining the required abilities suffices for epistemic desiderata satisfaction. In the other cases, our account allows us to specify which

understanding-related abilities support substantial desiderata satisfaction. This is a clear advantage of our account over less versatile approaches to understanding.

Example 2: Informed self-advocacy

Our ability clusters can also be used to specify the need for understanding described in existing work on desiderata in XAI. One example of that is our rephrasing of Rudin’s position as a requirement for the ability to design in the previous section. For another, more complex example in this vein, we can look at Vredenburg (2022)’s account of “informed self-advocacy”; a desideratum that even *requires* understanding, among other things, for substantial desiderata satisfaction. Her account relates to our proposal because she spells out informed self-advocacy in terms of abilities. Though these abilities do not readily fit with our clusters, our account allows us to specify exactly which understanding-related abilities are required for informed self-advocacy.

Vredenburg argues that informed self-advocacy relies on three abilities, two of which require insight into systems: The ability to “navigate systems of rules to achieve one’s goals” (Vredenburg, 2022, p. 213), which she calls *agency*, and the ability to hold decision-makers accountable if certain standards are not met, which she calls *accountability*.⁶⁶ These abilities are specified at a rather high level. We think the forms of understanding required to enable agency and accountability in specific contexts can be specified concretely via our abilities-based conceptualisation of understanding.

Recall our illustrative hiring scenario. To ensure her agency, it may be sufficient for April to possess the ability to *predict* whether she will be invited to an interview, that is, to anticipate the system’s behaviour given her application data. However, if April gets rejected, securing her agency would require the ability to *intervene* on the system’s input such that she will reach the interview in the next hiring round (e.g., by acquiring additional skills and updating her application accordingly). This could be supported by (ideally actionable, c.f., Karimi et al., 2021) counterfactual statements describing what would have been needed to be invited to an interview in the first hiring round (see also Wachter et al., 2017). The abilities of predicting and intervening enable April to navigate the system

⁶⁶ The third ability is the ability to represent one’s interests to decision-makers. As this ability is unrelated to the development and evaluation of an AI system (Vredenburg, 2022, p. 213), we will not discuss it here.

of rules that guides hiring decisions and therefore suffice to ensure April's agency in this context.

Let us move to accountability, which requires access to normative reasons for a decision and appropriate recourse should those reasons be unjust. If April gets rejected surprisingly (e.g., despite her excellent credentials), she might suspect that her rejection was based on discrimination. In order to ensure accountability, April needs the ability to assess whether the outcome "rejection" was appropriate. This could be supported by feature attribution methods revealing which information in April's application led to her rejection (see Speith, 2022b). If the revealed decision criteria are not self-explanatory, however, the explanation might need to be combined with normative reasons for their adequacy (Vredenburg, 2022, p. 217). This could require the hiring manager to be able to *explain* to April why she was rejected (Baum et al., 2022). With this combination of abilities in place, April can both detect flawed decisions and hold someone at the hiring company accountable for them. Overall, then, the understanding required to support April's informed self-advocacy can be spelled out in terms of the ability clusters we propose.

In sum, there is a close relationship between the understanding-related abilities stakeholders might possess and whether and how societal desiderata can be fulfilled. Specifying which abilities are necessary to fulfil a certain desideratum allows academics, advocates for ethical AI, and policy makers to be more precise about how, exactly, explainability can contribute to desiderata satisfaction. This can aid communication and collaboration between researchers, support the progress of the debate on XAI, and improve the development and evaluation of individual explainability approaches. Furthermore, we saw that abilities fit into existing work on the subject: We have shown that both Rudin's rejection of black-box models as well as Vredenburg's account of informed self-advocacy can be translated into demands for certain understanding-related abilities. Thus, we think that understanding-related abilities also meet the final criterion of a useful conceptualisation of understanding in XAI: abilities can be used to clarify the link between explainability, understanding, and the satisfaction of relevant societal desiderata.

2.7 Limitations and Objections

We have argued that abilities can serve as a useful conceptualisation of understanding as they meet the criteria we identified in Chapter 2.2: First, understanding-related abilities are theoretically motivated across relevant disciplines. Second, abilities are versatile enough to differentiate the various forms of understanding relevant to different XAI application contexts. Third, the abilities we propose are operationalisable. Finally, abilities can be used to identify necessary conditions for desiderata satisfaction and serve as a concrete description of the forms of understanding that are needed to meet a given desideratum in a specific context. As such, abilities can establish the often-assumed link between explainability, understanding, and satisfaction of stakeholder desiderata. Before concluding, we address three possible criticisms of our position.

Objection 1: Abilities are not necessary for understanding

Writing about *interpretability*, Krishnan (2020, p.491) rejects an ability-based view as “more plausibly a consequence of interpretability than it is a statement of that in which interpretability itself consists.” Thus, she judges pragmatic accounts of interpretability to be inadequate for contributing towards the objective of ethical AI. A similar objection could also be applied to understanding. Since contingent factors can prevent one’s cognitive state manifesting in practical abilities (Wilkenfeld, 2013), someone might possess understanding but not the abilities we describe. For instance, the ability to *explain* will normally require linguistic capabilities, and the ability to *intervene* presupposes (usually minimal) physical capacities. Similar requirements likely hold for other understanding-related abilities, too.

While abilities may not capture the concept of understanding completely, we think the more relevant question is whether they can specify the concept well enough to be useful. To that end, the previous sections have tried to show that a conceptualisation of understanding as abilities can have pragmatic value for XAI by fulfilling our four criteria. Furthermore, we believe that focusing on abilities can be of particular importance in XAI contexts, e.g., when morally weighty desiderata make it indispensable to ensure that the appropriate way of understanding was indeed afforded by an explainability approach; or when the collaboration of different stakeholders requires interaction with other agents

(e.g., by *explaining*) or with artifacts (e.g., by *intervening*). In such cases, these contingent factors that might not be constitutive of understanding are nevertheless required for desiderata satisfaction.

Objection 2: Abilities are not sufficient for understanding

Another worry is that abilities may be present in the absence of understanding (see also Wilkenfeld, 2013). For instance, someone could demonstrate an ability to *explain* by parroting an explanation without truly grasping it. This scenario could occur in contexts where a stakeholder offloads the cognitive work involved in understanding to a reliable explainability approach. Similarly, for systems that can only provide a small range of possible outputs, someone might be able to predict the system's behaviour by mere luck, without possessing any understanding. Similar scenarios are conceivable for the other abilities we suggest.

However, it is doubtful whether the abilities as we have described them in Chapter 2.4 are indeed present in the above cases. Mere parroting arguably does not count as explaining, even if it appears otherwise; and lucky guessing seems to be fundamentally different from predicting. Instead of rejecting abilities as unreliable, researchers interested in measuring understanding need to be careful when operationalising it. For instance, parroting and lucky guessing could be detected by demanding participants to reformulate an explanation or to make predictions on a range of counterfactual cases. A requirement for valid and robust empirical evaluations should ensure that stakeholders reliably possess the abilities in question. Such a requirement should also rule out subjective measures (e.g., self-reports) which cannot discriminate between a stakeholder's sense of understanding and the practical abilities associated with it.

Robust assessment of abilities is especially important in high-stakes contexts, where misjudgements could lead to harmful outcomes, and in what Bordt et al. (2022) call *adversarial explanation contexts*, where the parties involved in building, explaining, and using an AI system have conflicting interests. In such contexts, the explanation provider may be incentivised to furnish the explanation recipient with misleading explanations that induce false confidence in one's abilities or a false sense of understanding (i.e., one not accompanied by the relevant abilities; Bordt et al., 2022).

However, even if a reliable method indicates that a person possesses relevant abilities, this may still be insufficient for a demanding, philosophical account of understanding (for an overview of possibilities, see Baumberger et al., 2017; Baumberger, 2019). To someone who holds this kind of position, specific measures of abilities can only be fallible proxies for understanding; they suggest that it is present, but do not guarantee it. Though we grant that this may be the case, our goal was not to provide a metaphysical definition of understanding, but a useful conceptualisation that is able to fulfil the role that is commonly attributed to understanding in XAI. We think abilities are the most suitable candidate for this role.

Objection 3: Characterising understanding is futile

In light of the difficulties that arise when one tries to specify the relationship between understanding and abilities, one might question whether these efforts are worthwhile. Krishnan (2020) argues that research addressing ethical issues arising from opaque AI algorithms need not focus on psychological constructs to achieve its goal of satisfying societal desiderata. In Krishnan's view, the challenges inherent to specifying nebulous terms like explainability and understanding render them more problematic than useful. Instead of trying to characterise these terms, one should directly address the actual aim, that is, desiderata satisfaction.

We believe this argument is premature in rejecting a pragmatic approach to dealing with concepts like understanding. We agree with Krishnan that there is no *one* way in which systems can be explainable, nor *one* form of understanding which will help satisfy all relevant desiderata. We disagree, however, that this is grounds to focus solely on desiderata. Instead, we believe our arguments have shown that conceptualising understanding via abilities is a fruitful way to specify forms of understanding which can build a bridge between explainability approaches and desiderata. We believe that this approach is preferable to looking directly at desiderata for two reasons.

First, clusters of abilities allow for the identification of commonalities between the epistemic requirements for different desiderata.⁶⁷ Langer et al. (2021c) and Chazette et

⁶⁷ One could argue that our position is closer to Krishnan's than it might superficially appear, as we identify only the epistemic requirements of desiderata satisfaction rather than describing the metaphysical nature

al. (2021) identify numerous desiderata that are discussed within the XAI literature. For every desideratum, the lack of understanding of AI systems plays a role. While this cannot be addressed in the same way for every desideratum, there are recurring problems that our pragmatic conceptualisation of understanding can capture. Second, an independent characterisation of understanding allows for greater specificity about the role of explainability for desiderata satisfaction. In some cases, we rely on explainability to guarantee epistemic desiderata satisfaction, that is, to gain epistemic access to the properties of a system, even when technical requirements on AI systems are fully satisfied (see also Langer et al., 2021c). In other cases, explainability is expected to contribute to substantial desiderata satisfaction, but it remains unclear how exactly. For example, it is still unclear whether and how explanations can calibrate trust in a system (Kästner et al., 2021), and it is ethically preferable only to use explanations to increase trust when it routes through an adequate ability to assess the trustworthiness of the system (Schlicker & Langer, 2021). For reasoning about both kinds of cases, a concrete conceptualisation of understanding that is independent of desiderata satisfaction is required. Therefore, we consider our approach to be valuable even for those primarily interested in desiderata satisfaction.

2.8 Conclusion

In this paper, we proposed a conceptualisation of understanding useful for XAI endeavours: *abilities*. More specifically, we suggested that someone who understands some aspect(s) of an AI system typically possesses one or several understanding-related abilities (viz., the abilities to recognise, to assess, to predict, to explain, to intervene, and to design). Ceteris paribus, a person who possesses an ability to a higher degree than another person also has a higher degree of understanding. Furthermore, as indicated by the hierarchical organisation of our ability clusters, the respective abilities tend to correspond to different degrees of understanding. To support our proposal, we showed that abilities are motivated by interdisciplinary research, versatile, operationalisable, and

of understanding. However, given the omnipresence of “understanding” in the XAI debate, we consider it to be more useful to clarify different uses of understanding as opposed to eliminating talk of understanding. We thank an anonymous reviewer for pressing this point.

can be used to clarify the link between understanding and desiderata satisfaction. As such, our approach can benefit various stakeholders involved in XAI, both inside and outside computer science. For example, abilities can be used to precisely specify understanding-related requirements for the use and deployment of AI systems. In turn, having concrete requirements spelled out can help developers design new explainability approaches with a clear view of what they need to achieve.

Overall, the abilities-based conceptualisation of understanding we propose can serve as a common language that facilitates communication and collaboration between different stakeholders involved in XAI. However, as a next step for such a conceptualisation to effectively address challenges in XAI, research needs to investigate which experimental paradigms best operationalise understanding-related abilities. Furthermore, work exploring how particular abilities relate to specific desiderata would strengthen the theoretical basis on which to deploy explainability approaches in real-world contexts. Finally, our conceptualisation of understanding in XAI contexts could inform a formal definition of this concept that takes additional philosophical and psychological debates on understanding into account. We believe that the clusters we have proposed will prove useful for these and related projects.

2.9 References

- Allahyari, H., & Lavesson, N. (2011). User-oriented assessment of classification model understandability. In Kofod-Petersen, A., Heintz, F., & Langseth, H. (Eds.), *Proceedings of the 11th Scandinavian Conference on Artificial Intelligence (SCAI 2011)*, Frontiers in Artificial Intelligence and Applications, Vol. 227. IOS Press, pp. 11–19, <https://doi.org/10.3233/978-1-60750-754-3-11>
- Alqaraawi, A., Schuessler, M., Weiß, P., et al. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In Paternò, F., Oliver, N., Conati, C., et al. (Eds.), *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI 2020)*. Association for Computing Machinery, pp. 275–285, <https://doi.org/10.1145/3377325.3377519>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>

- Baum, K., Mantel, S., Schmidt, E., et al. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1), 12. <https://doi.org/10.1007/s13347-022-00510-w>
- Baumberger, C. (2014). Types of understanding: Their nature and their relation to knowledge. *Conceptus*, 40(98), 67–88. <https://doi.org/10.1515/cpt-2014-0002>
- Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science*, 50(3), 367–388. <https://doi.org/10.1007/s10838-019-09474-6>
- Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). Routledge.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., et al. (1965). Taxonomy of educational objectives. *Handbook 1: Cognitive domain*. McKay.
- Bordt, S., Finck, M., Raidl, E., et al. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In Isbell, C., Lazar, S., Oh, A., et al. (Eds.), *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*. Association for Computing Machinery, pp. 891–905, <https://doi.org/10.1145/3531146.3533153>
- Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, 81(6), 1211–1241. <https://doi.org/10.1007/s10670-015-9791-5>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>
- Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, 5(3). <https://doi.org/10.23915/distill.00024>
- Carnap, R. (1962). *Logical foundations of probability*. University of Chicago Press.
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 1–34. <https://doi.org/10.3390/electronics8080832>
- Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring explainability: A definition, a model, and a knowledge catalogue. In Cleland-Huang, J., Moreira, A., Schneider, K., et al. (Eds.), *Proceedings of the 29th IEEE International Requirements Engineering Conference (RE 2021)*. IEEE, pp. 197–208, <https://doi.org/10.1109/RE51729.2021.00025>
- Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*, 71(1), 287–319. <https://doi.org/10.1093/bjps/axy015>
- Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In Schneider, K., Dalpiaz, F., & Horkoff, J. (Eds.), *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops (REW 2023)*. IEEE, pp. 316–324, <https://doi.org/10.1109/REW57809.2023.00060>
- Deck, L., Schoeffer, J., De-Arteaga, M., et al. (2024). A critical survey on fairness benefits of XAI. In F. Steibel, M. Young & R. Baeza-Yates (Eds.), *Proceedings of the 7th ACM*

Conference on Fairness, Accountability, and Transparency (FAccT 2024). Association for Computing Machinery. <http://arxiv.org/abs/2310.13007>

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7), 1829–1850.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64. <https://doi.org/10.1518/001872095779049543>

Franconeri, S. L., Padilla, L. M., Shah, P., et al. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, 22(3), 110–161. <https://doi.org/10.1177/15291006211051956>

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4), 111–117. <https://doi.org/10.1215/07402775-3813015>

Grimm, S. R. (2011). Understanding. In S. Bernecker & D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 84–94). <https://doi.org/10.4324/9780203839065.ch9>

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4), 889–911. <https://doi.org/10.1093/bjps/axi148>

Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly*, 58(3), 269–290. <https://doi.org/10.2307/48616060>

Hickman, L., Bosch, N., Ng, V., et al. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8), 1323–1351. <https://doi.org/10.1037/apl0000695>

Hills, A. (2016). Understanding why. *Noûs*, 50(4), 661–688. <https://doi.org/10.1111/nous.12092>

Hoffman, R. R., Mueller, S. T., Klein, G., et al. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv*. <https://doi.org/10.48550/arXiv.1812.04608>

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In Boyd, D., & Morgenstern, J. H. (Eds.), *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency* (FAT* 2019). Association for Computing Machinery, pp. 49–58, <https://doi.org/10.1145/3287560.3287600>

Huysmans, J., Dejaeger, K., Mues, C., et al. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1), 141–154. <https://doi.org/10.1016/j.dss.2010.12.003>

Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In Elish, M. C., Isaac, W., & Zemel, R. S. (Eds.), *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2021). Association for Computing Machinery, New York, NY, USA, pp 353–362, <https://doi.org/10.1145/3442188.3445899>

Kästner, L., Langer, M., Lazar, V., et al. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. In Yue, T., & Mirakhorli, M. (Eds.), *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops* (REW 2021). IEEE, pp. 169–175, <https://doi.org/10.1109/REW53955.2021.00031>

Khalifa, K. (2013). Understanding, grasping and luck. *Episteme*, 10(1), 1–17. <https://doi.org/10.1017/epi.2013.6>

- Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*, 84, 46–56. <https://doi.org/10.1016/j.shpsa.2020.08.003>
- Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3), 487–502. <https://doi.org/10.1007/s13347-019-00372-9>
- Kvanvig, J. (2009). The value of understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 95–111). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199231188.003.0005>
- Langer, M., & König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review*, 33(1), 100881. <https://doi.org/10.1016/j.hrmr.2021.100881>
- Langer, M., Baum, K., Hartmann, K., et al. (2021a). Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In Yue, T., & Mirakhorli, M. (Eds.), *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops (REW 2021)*. IEEE, pp. 164–168, <https://doi.org/10.1109/REW53955.2021.00030>
- Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology*, 36(5), 751–769. <https://doi.org/10.1007/s10869-020-09711-6>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021c). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Lapuschkin, S., Binder, A., Montavon, G., et al. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In Tuytelaars, T., Li, F. F., Bajcsy, R., et al. (Eds.), *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. IEEE, pp. 2912–2920, <https://doi.org/10.1109/CVPR.2016.318>
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99(2), 167–204. <https://doi.org/10.1016/j.cognition.2004.12.009>
- Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023). Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy. *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 337–342. <https://doi.org/10.1109/REW57809.2023.00063>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>

- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum. Or: How I learnt to stop worrying and love the social and behavioural sciences. In Aha, D. W., Darrell, T., Pazzani, M., et al. (Eds.), *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence* (XAI 2017). IJCAI, pp. 36–42, arXiv:1712.00547
- Newman, M. (2017). An evidentialist account of explanatory understanding. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 190–211). Routledge.
- Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Piltaver, R., Luštreka, M., Gams, M., & Martinčić-Ipšić, S. (2014). Comprehensibility of classification trees-Survey design. In *Proceedings of 17th International Multiconference Information Society* (IS 2014). Information Society, pp. 70–73
- Potochnik, A. (2010). Levels of explanation reconceived. *Philosophy of Science*, 77(1), 59–72. <https://doi.org/10.1086/650208>
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., et al. (2021). Manipulating and measuring model interpretability. In Kitamura, Y., Quigley, A., Isbister, K., et al (Eds.), *Proceedings of the 39th ACM Conference on Human Factors in Computing Systems* (CHI 2021). Association for Computing Machinery, pp. 237:1–237:52. <https://doi.org/10.1145/3411764.3445315>
- de Regt, H. W. (2015). Scientific understanding: Truth or dare? *Synthese*, 192(12), 3781–3797. <https://doi.org/10.1007/s11229-014-0538-7>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., et al. (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2016). Association for Computing Machinery, pp 1135–1144, <https://doi.org/10.1145/2939672.2939778>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In McIlraith, S. A., & Weinberger, K. Q. (Eds.), *Proceedings of the 32nd AAAI Conference on Artificial Intelligence* (AAAI 2018), *the 30th Innovative Applications of Artificial Intelligence Conference* (IAAI 2018), and *the 8th AAAI Symposium on Educational Advances in Artificial Intelligence* (EAAI 2018). AAAI Press, pp. 1527–1535, <https://doi.org/10.1609/aaai.v32i1.11491>
- Riggs, W. D. (2003). Intellectual virtue: Perspectives from ethics and epistemology. In M. DePaul & L. Zagzebski (Eds.), *Understanding ‘virtue’ and the virtue of understanding* (pp. 203–226). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199252732.003.0010>
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In Schneegass, S., Pfleging, B., & Kern, D. (Eds.), *Proceedings of Mensch und Computer (MuC 2021)*. Association for Computing Machinery, pp. 325–329, <https://doi.org/10.1145/3473856.3474018>.
- Speith, T. (2022a). How to evaluate explainability: A case for three criteria. In Knauss, E., Mussbacher, G., Arora, C., et al. (Eds.), *Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops (REW 2022)*. IEEE, pp. 92–97, <https://doi.org/10.1109/REW56159.2022.00024>
- Speith, T. (2022b). A review of taxonomies of explainable artificial intelligence (XAI) methods. In Isbell, C., Lazar, S., Oh, A., et al. (Eds.), *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*. Association for Computing Machinery, pp. 2239–2250, <https://doi.org/10.1145/3531146.3534639>
- Stock, P., Cissé, M. (2018). ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In Ferrari, V., Hebert, M., Sminchisescu, C., et al. (Eds.), *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018)*, *Lecture Notes in Computer Science*, Vol. 11210. Springer International Publishing, pp. 504–519, https://doi.org/10.1007/978-3-030-01231-1_31
- Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press. <https://doi.org/10.2307/j.ctv1dv0tnw>
- Thórisson, K. R., Kremelberg, D., Steunebrink, B. R., et al. (2016). About understanding. In Steunebrink, B., Wang, P., & Goertzel, B. (Eds.), *Proceedings of the 9th International Conference on Artificial General Intelligence (AGI 2016)*, *Lecture Notes in Computer Science*, Vol. 9782. Springer International Publishing, pp. 106–117, https://doi.org/10.1007/978-3-319-41649-6_11.
- Tullio, J., Dey, A. K., Chalecki, J., et al. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In Rosson, M. B., & Gilmore, D. J. (Eds.), *Proceedings of the 25th ACM Conference on Human Factors in Computing Systems (CHI 2007)*. Association for Computing Machinery, pp. 31–40, <https://doi.org/10.1145/1240624.1240630>.
- Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2015). Goals affect the perceived quality of explanations. In Noelle DC, Dale R, Warlaumont AS, et al (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*. Cognitive Science Society, pp. 2469–2474, <https://cogsci.mindmodeling.org/2015/papers/0424/paper0424.pdf>
- Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review*, 24(5), 1436–1450. <https://doi.org/10.3758/s13423-017-1275-y>
- Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3), 615–661. <https://doi.org/10.3390/make3030032>

Vredenburg, K. (2022). The right to explanation. *Journal of Political Philosophy*, 30(2), 209–229. <https://doi.org/10.1111/jopp.12262>

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. <https://doi.org/10.2139/ssrn.3063289>

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, 190(6), 997–1016. <https://doi.org/10.1007/s11229-011-0055-x>

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalisation: Evidence from category learning. *Cognitive Science*, 34(5), 776–806. <https://doi.org/10.1111/j.1551-6709.2010.01113.x>

Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In Ohlsson S, Catrambone R (Eds.), *Proceedings of the 32th Annual Meeting of the Cognitive Science Society (CogSci 2010)*. Cognitive Science Society, pp. 2906–2911

Chapter 3:

Explaining AI Through Mechanistic Interpretability

This chapter was published as “Kästner, L., & Crook, B. (2024). Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, 14(4), 52.” Slight changes have been made to fit the dissertation format.

3.0 Abstract

Recent work in explainable artificial intelligence (XAI) attempts to render opaque AI systems understandable through a divide-and-conquer strategy. However, this fails to illuminate *how trained AI systems work as a whole*. Precisely this kind of functional understanding is needed, though, to satisfy important societal desiderata such as safety. To remedy this situation, we argue, AI researchers should seek *mechanistic interpretability*, viz. apply coordinated discovery strategies familiar from the life sciences to uncover the functional organisation of complex AI systems. Additionally, theorists should accommodate for the unique costs and benefits of such strategies in their portrayals of XAI research.

3.1 Introduction

Over the past decade, the term “AI” has increasingly become a synonym for deep artificial neural networks (ANNs) trained with machine learning (ML) algorithms. These ANNs are often complex and opaque, approximating target functions through the mutual contribution of millions or even billions of parameters with values learned during an automated training process (Baraniuk et al., 2020; Chollet, 2021; Russell & Norvig, 2020). On the one hand, this setup allows ANNs to exhibit flexible and expressive behaviour, developing sophisticated ways of representing and processing information. On the other hand, these systems can be surprisingly brittle and prone to unpredictable and catastrophic failure (Fawzi et al., 2018; Raghu et al., 2017). Since the functional

organisation of ANNs is both complex and machine-learned, *how* their internal structure implements the mapping from inputs to outputs often remains unknown. Despite this, AI systems based on ANNs are increasingly used to support or take over human tasks – from making music recommendations to medical decisions (Huang et al., 2020; Jugovac & Jannach, 2017). Especially in high-stakes⁶⁸ circumstances, regulators, philosophers and AI researchers increasingly call for making AI systems *explainable* (Burrell, 2016; Zednik, 2021).

To address this need, a research field has developed under the heading of explainable AI (XAI). But far from being unified, research under this umbrella tackles a range of distinct, though interrelated issues: Computer scientists take on the technical challenge of developing computational methods to produce explanations of AI systems (Guidotti et al., 2018; Ribeiro et al., 2016; Wachter et al., 2017). Philosophers attempt to characterise relevant societal desiderata such as safety, trustworthiness, reliability, and fairness (Durán & Jongsma, 2021; Langer, Oster, et al., 2021; Páez, 2019). And psychologists grapple with the problem of quantifying epistemic outcomes like *understanding* (Langer, Baum, et al., 2021; Sloman & Rabb, 2016). In the face of this complexity, scholars have tried to coordinate these different strands of research, highlight how they might be mutually advantageous, and assess the extent to which (and at which costs) XAI can achieve its goals. To this end, XAI researchers have developed taxonomies of explainability methods (Nunes & Jannach, 2017; Speith, 2022), offered conceptual models of the so-called explainability problem (Langer, Oster, et al., 2021; Zednik, 2021), and proposed simplifying unifications (Fleisher, 2022; Nyrup & Robinson, 2022). We applaud these contributions, and agree that AI explainability is a multifaceted issue which resists any one-size-fits-all solution (cf., Langer, Oster, et al., 2021; Zednik, 2021). However, the picture of XAI research that is typically presented is one in which computational methods deliver explanations for specific stakeholders in particular contexts. We think this is crucially incomplete: we must also consider the potential of

⁶⁸ We refrain from committing to any particular definition of ‘high-stakes’. See the European Commission’s Artificial Intelligence Act (Artificial Intelligence Act, 2024) for extended discussion of what constitutes a ‘high-risk’ application of AI.

coordinated research strategies aiming to uncover the functional organisation of trained AI systems.

While research to this end is currently gaining momentum among researchers working in industry (e.g., Cammarata et al., 2020; Olah et al., 2018), attention to it is still limited within the academic sphere (but see, e.g., Bau et al., 2017; Geiger et al., 2022, 2023). Contemporary XAI effectively pursues a *divide-and-conquer strategy* by focusing on how individual methods might deliver explainability in specific contexts. While this strategy can successfully highlight which features were influential for a given outcome, or how input features would need to be changed to obtain a different decision, this is only a small part of the story if we are seeking to explain and understand how AI systems work. Indeed, what we *should* aim for to ensure that AI systems fulfil desiderata commonly demanded of AI systems by society – such as safety – goes far beyond specific contexts: We need generalisable insights about *how* the systems in question *work as a whole*; for only such insights about the functional organisation that elicits behavioural patterns and dispositions will allow us to anticipate how systems might respond to novel inputs and how they might behave when exposed to yet unexplored contexts. Importantly, what is of interest here is not the architecture that an AI system’s programmers have specified prior to training.⁶⁹ Rather, we are interested in the learned structure that emerges through the automated training procedures a system undergoes after its initial specification; hence it is also referred to as *emergent structure* (Manning et al., 2020). Crucially, we argue, these insights cannot usually be extrapolated from applying individual XAI methods tailored to specific contexts – we would simply be missing the forest for the trees.

We propose that to remedy this situation XAI researchers should take a *mechanistic interpretability (MI) approach* for complex AI systems. In fact, research in line with the approach we suggest is already being pursued by a small community of researchers (Chapter 3.3.2). It starts from the premise that once AI systems become sufficiently complex, they are best investigated and explained through the same lens as biological organisms (rather than being treated as technical artefacts). Thus, practitioners should

⁶⁹ For ANNs, the number of layers, number of units (neurons) per layer, activation functions of units, connectivity of units, loss function, and learning algorithm (including hyperparameters like batch size and learning rate) are all typically pre-programmed.

seek to characterise AI systems in terms of their functional organisation (the organised activities of their functionally relevant components).⁷⁰ This requires the application of coordinated discovery strategies familiar from life sciences, such as pattern recognition, functional decomposition, localisation, and systematic experimental manipulations. As such, MI research may be significantly more resource-intensive (both in terms of time and in terms of labour) than developing specific XAI methods. In return for this investment, we gain a deeper and more holistic understanding of how trained AI systems work. We argue that the outcome of effective MI research affords greater predictivity and, by enabling surgical interventions, greater control of system behaviour. Thus, we claim, MI enables us to meet important societal desiderata including safety. Given these distinctive costs and benefits, and the fact that the MI approach has been successfully employed by a small community of researchers, we think it is crucial for XAI theorists (as well as philosophers of science more generally) to accommodate for this research strategy in their analyses of the field. For only by having the full range of strategies for explaining complex AI systems in view can we make appropriate choices about which research to pursue, and thereby ensure that crucial desiderata will actually be satisfied when employing opaque AI systems in society.

We shall proceed as follows: In Chapter 3.2, we briefly discuss achievements and limitations of contemporary XAI research. In Chapter 3.3, we propose that rather than adopting a divide-and-conquer strategy, XAI researchers should seek explainability through mechanistic interpretability. We illustrate this approach by examining the case of Distill. In Chapter 3.4, we discuss possible objections to our proposal. In Chapter 3.5, we conclude.

3.2 Contemporary XAI: The Divide-and-Conquer Strategy

For current purposes we focus on ML-based AI systems (with ANNs being a paradigm case) that are opaque due to the scale and complexity of their learned structure (Burrell,

⁷⁰ Functionally relevant components can be any unit or structure that serves to achieve a specific function within a system; this is explicitly not limited to pre-specified entities like neurons but may include complex structures such as, e.g., circuits (see Chapter 3.3.1) or distributed representations.

2016). Following Humphreys (2009), we characterise opacity as a lack of knowledge about a system’s epistemically relevant elements (EREs) (see also Zednik, 2021).⁷¹ The term ERE is deliberately non-specific. It captures any robust patterns which underlie or maintain a system’s behaviour and are relevant to the epistemic goals of an agent. Thus, a system’s opacity is relative both to an agent’s (e.g., a company, AI user, or developer) interests and their knowledge about the system at a given time. Against this backdrop, we assume that any information eliminating opacity (by uncovering EREs) can function as an explanation (cf., Nyrup & Robinson, 2022; Zednik, 2021). We take making AI systems explainable by reducing their opacity to be the goal of XAI.⁷²

Contemporary XAI research tries to achieve this goal through technical means. Specifically, it aims to develop algorithmic procedures – XAI *methods* – that generate explanatory information about AI systems. Hence, XAI is very much “in the business of developing analytic techniques with which to render opaque computing systems transparent” (Zednik, 2021, p. 285; see Mittelstadt et al., 2019; Rudin, 2019 for similar claims). Since the kind of explanatory information needed to eliminate opacity depends on numerous factors (Kirsch, 2017; Langer et al., 2021), XAI researchers have developed a range of methods with different properties (e.g., scope, detail, format) (Guidotti et al., 2018; Molnar, 2022; Speith, 2022). Further work has addressed which methods are best suited for concrete scenarios in which stakeholders interact with AI systems (Barredo Arrieta et al., 2020; Belle & Papantonis, 2021). This approach assumes that rendering AI systems explainable requires i) developing a range of XAI methods, ii) identifying specific explainability contexts, and iii) mapping XAI methods to these contexts. The result is a *divide-and-conquer strategy*: XAI research seeks, in any given case, to provide “the most appropriate explanation for a specific ML solution in a given context on a given task” (Zhou et al., 2021, p. 2; see also Fleisher, 2022, p. 12).

⁷¹ Notice that while we adopt Humphreys’ framing in terms of *knowledge* of EREs, this can be adapted to an account centred on *understanding* by defining opacity as the lack of a *grasp* of EREs. Grasping an ERE involves making use of it to perform inferences, take decisions, or complete downstream tasks (Keil, 2019; Strevens, 2013).

⁷² Achieving this outcome does not require comprehensive explanation of *all* AI systems. For systems that operate in low-stakes environments, fewer elements of the system may be deemed epistemically relevant.

The divide-and-conquer strategy has its merits. It naturally accommodates the insight that there cannot be a one-size-fits all XAI approach, and that different stakeholders, contexts and AI systems pose different constraints on feasible XAI methods. Further, the divide-and-conquer strategy has been and continues to be productive, both in terms of novel methods and conceptual analysis (e.g., Barredo Arrieta et al., 2020; Belle & Papantonis, 2021; Köhl et al., 2019; Langer et al., 2021; Ribeiro et al., 2016; Sokol & Flach, 2020; Zhou et al., 2021). Indeed, contemporary XAI research successfully produces tools enabling stakeholders to understand narrow aspects of AI systems (Belle & Papantonis, 2021; Guidotti et al., 2018; Lapuschkin et al., 2019; Molnar, 2022; Ribeiro et al., 2016; Ribeiro et al., 2018; Wachter et al., 2017). One prominent example is *Local Interpretable Model-Agnostic Explanations* (LIME) (Ribeiro et al., 2016). This technique, applicable to any AI classification system, produces linear surrogate models which highlight the importance of particular features for a given classification. LIME has been shown to improve stakeholders' ability to assess the quality of rival classifiers and, as such, may prove an adequate solution in specific contexts where such skills are required. Another success story for the divide-and-conquer strategy is research on *counterfactual explanations* and *algorithmic recourse* (Karimi et al., 2022; Wachter et al., 2017). This family of XAI methods can inform subjects of algorithmic decisions about which alterations to the input data would have resulted in a desired outcome. Such methods are crucial to enabling agency in an increasingly algorithmic world (Vredenburg, 2022).

Despite these successes, however, the *divide-and-conquer strategy* is also limited. Neither LIME nor counterfactual explanations can provide a comprehensive understanding of *how the systems* under investigation *work as a whole*. They do not tell us *how exactly* the internal structure of the model maps inputs to outputs, neither do they allow us to predict how a system will behave in *novel contexts*. Indeed, in an influential paper, Rudin criticises the enterprise of XAI for this reason, asserting that XAI methods “do not provide enough detail to understand what the black box is doing” (2019, p. 208). Similarly, Freiesleben (2024) evaluates two specific XAI methods (*feature visualisation* and *network dissection*) and concludes that neither are sufficient to support inferences about the functional role of subparts of model structure. Importantly, this information cannot be inferred even by independent application of various XAI method outputs. What

is needed instead is systematic investigation that yields generalisable insights about systems' overall functional organisation.

The need for such insights is expressed vividly in current debates about AI regulation in politics and society. Take the European Commission's proposed regulatory framework for AI systems (Artificial Intelligence Act, 2024). Article 14 states that adequate human oversight in high-risk scenarios demands that an overseer should "properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, also in view of detecting and addressing anomalies, dysfunctions and unexpected performance" and be able "to intervene on the operation of the high-risk AI system" (Artificial Intelligence Act, 2024, art. 14). With regards to both requirements, a detailed (and generalisable) understanding of a trained systems' overall functional organisation is crucial. For illustration, consider the phenomenon of *typographic adversarial examples* in the image model *CLIP* (Goh et al., 2021). CLIP was trained to predict which text was paired with a given image on the internet. However, researchers found that CLIP was easily fooled by simply sticking a written label (e.g., "phone") on an object (e.g., an apple), causing the image to be classified according to the written label rather than the object (e.g., as phone rather than apple). This behaviour was not foreseeable from ordinary reliability testing. Neither does highlighting the paper label as relevant to the classification output explain *why* CLIP relied on the labels rather than the objects in the images for classification. However, through systematic investigation into the model, researchers revealed high-level neurons in CLIP which are sensitive to both images and text (Goh et al., 2021). Understanding this feature of the model's functional organisation made it easy to explain and predict CLIP's behaviour when faced with written labels in images.

Luckily, CLIP is not in use in high-stakes domains. However, consider the dangers of deploying a system with similar vulnerabilities to, e.g., scan for weapons at an airport. Here, as in many other cases where AI is deployed in the real world, safety is among the chief desiderata. It is listed among the first objectives in the EU AI Act (Artificial

Intelligence Act, 2024, art. 1).⁷³ One way to ensure safety and prevent harm is to subject AI systems to rigorous testing before deploying them in high-stakes situations (Durán & Jongsma, 2021). However, doing this exhaustively will often be infeasible due to the dimensionality of the input space (Hacker et al., 2023; Keogh & Mueen, 2017). This makes it difficult to rule out the possibility of unpredictable failure when systems encounter novel inputs (Amodei et al., 2016; Hendrycks et al., 2018, 2023; Wei et al., 2023). A necessary complementary strategy, we think, is to characterise how the systems in question function in terms of how relevant components work together to elicit the behaviour we observe, i.e., MI. Based on such characterisations, we can anticipate and interpret system behaviour even for novel situations, helping us to ensure system reliability and safety (Meyes et al., 2019). The case of CLIP goes to show that insights into a system’s functional organisation can help us understand a system’s capabilities and limitations and avoid unforeseen failures.

Another benefit of MI is enabling more precise and informed scientific communication about the nature and properties of opaque AI systems. The high-profile release of highly performant models such as ChatGPT has led to a surge of debate about these systems (M. Mitchell & Krakauer, 2023; Pavlick, 2023). However, whether or not it is reasonable to describe large language models (LLMs) such as ChatGPT as possessing knowledge (Lam, 2022), implementing symbolic reasoning (Pavlick, 2023), or representing the world (Li et al., 2023) arguably depends upon empirical questions about the internal structure that implements the model’s behaviour (Millière & Buckner, 2024). In other words, an accurate and scientifically defensible description of AI systems is contingent on the characterisation of EREs that may only be achieved through MI. We discuss the advantages of MI further in Chapter 3.3.3.

Importantly, the issue we are getting at here is a principled one, not a criticism of any particular XAI method. We acknowledge that certain methods are sufficient for explaining particular instances of system behaviour. But when deploying AI systems in the real world, fulfilling important societal desiderata requires a kind of *holistic explainability* that

⁷³ While there is no domain-general definition of what it means for an AI system to be safe, safety is typically associated with the avoidance of both physical and psychological harm to human beings (Steimers & Schneider, 2022).

provides generalisable insights into how a system's functional organisation elicits its behaviours or outputs.⁷⁴

For clarity, we stress that the holistic explanations MI aims for are not synonymous with *global* explanations. In the XAI literature, *global* is an adjective ascribed to XAI methods which produce explanatory information about whole models (Molnar, 2022; Speith, 2022); it is contrasted with *local* methods which produce explanatory information about individual model inferences (e.g., one prediction or classification). *Holistic*, in contrast to both of these terms, is an adjective we use to refer to explanations (such explanations typically involve diagrams and accompanying text) which capture the *entire functional structure* of trained models (as MI aims to do). In principle, an idealised global XAI method *could* produce a holistic explanation. However, in practice, this is clearly not the case. For example, Molnar (2022) categorises *partial dependence plots* as global methods, but plotting the marginal effect of particular features on model outputs is clearly a far cry from providing a holistic explanation of a trained model's functional structure.

Ultimately, we claim that the project of developing holistic explanations of sufficiently complex AI systems will *never* be adequately achieved through employing specific XAI methods in particular contexts. Instead, we propose, it can be achieved by *mechanistic interpretability* (MI). That is, by developing (human-interpretable) *mechanistic* explanations of systems' functional organisation. So far, this kind of project has been underappreciated by both philosophers and computer scientists.

3.3 Augmenting Contemporary XAI Through Mechanistic Interpretability

We suggest taking a *mechanistic interpretability (MI) approach* to complex AI systems that starts from the following premise: once AI systems become sufficiently complex, they are best investigated and explained through the same lens as biological organisms

⁷⁴ We use the term 'holistic' to signal a focus on how AI systems' functional organisation implements their behaviour. As we will show, this need not be opposed to reductionist methodologies (see also Burnston, 2021).

rather than being treated merely as technical artefacts (cf., Eden, 2007). Taking inspiration from successful scientific inquiry in the life sciences, we propose that XAI practitioners should apply coordinated discovery strategies (such as pattern recognition, functional decomposition, localisation, and systematic experimental manipulations) to characterise AI systems in terms of their functional organisation, i.e., the organised activities of their functionally relevant components. At the same time, theorists should accommodate the unique costs and benefits of this research strategy in their portrayals of XAI research. With the full range of strategies for explaining AI systems in view, we can make appropriate choices about how to ensure that crucial desiderata will be satisfied when deploying trained AI systems in society.

3.3.1 Mechanistic interpretability

If we want to explain how AI systems work *as a whole*, we are essentially interested in their *functional organisation* or structure. That is, we seek to understand what system properties support their behaviour and how system functions are implemented by the orchestrated interactions of relevant component parts. This kind of project is neither new nor unique to XAI research.⁷⁵ And it is highly familiar from recent discussions on how to explain biological systems mechanistically (see Bechtel, 2009; Bechtel & Richardson, 1993; Craver, 2001; Craver & Darden, 2013). When mechanistic inquiry delivers explanations of sufficiently high quality, it affords improved prediction and control of target systems (Baetu, 2011; Craver, 2007; Howick et al., 2010; Winning & Bechtel, 2018; Woodward, 2017; Zou et al., 2023). Therefore, we suggest adopting strategies familiar from mechanistic inquiry in the life sciences to develop mechanistic explanations of the capabilities, limitations, and behaviours of opaque AI systems.

Recall that ANNs acquire their functional organisation through automated training procedures. After training, components within the system (the EREs of mechanistic explanations) will adopt specialised roles that programmers do not usually anticipate. These components will be organised in ways allowing them to work together to elicit the behaviours we observe (Manning et al., 2020; Richards et al., 2019). The task for

⁷⁵ The idea that certain aspects of computer science should be treated as empirical inquiry (as opposed to a branch of mathematics) goes back at least as far as Newell and Simon (1976), see Eden (2007) for discussion.

researchers seeking MI will be to uncover the relevant components along with their organisation. This task may not be simple, and it may seem impossible to map functions onto discrete parts of an ANN at first sight (see Chapter 3.4 for further discussion). In fact, identifying relevant components will often go beyond investigating the functional roles of pre-individuated ANN structures like neurons and layers (Bau et al., 2017). Frequently, it may involve characterising more exotic and distributed structures like circuits (see Chapter 3.3.2), high-level representations (Zou et al., 2023), and representational subspaces (see Elhage et al., 2021).

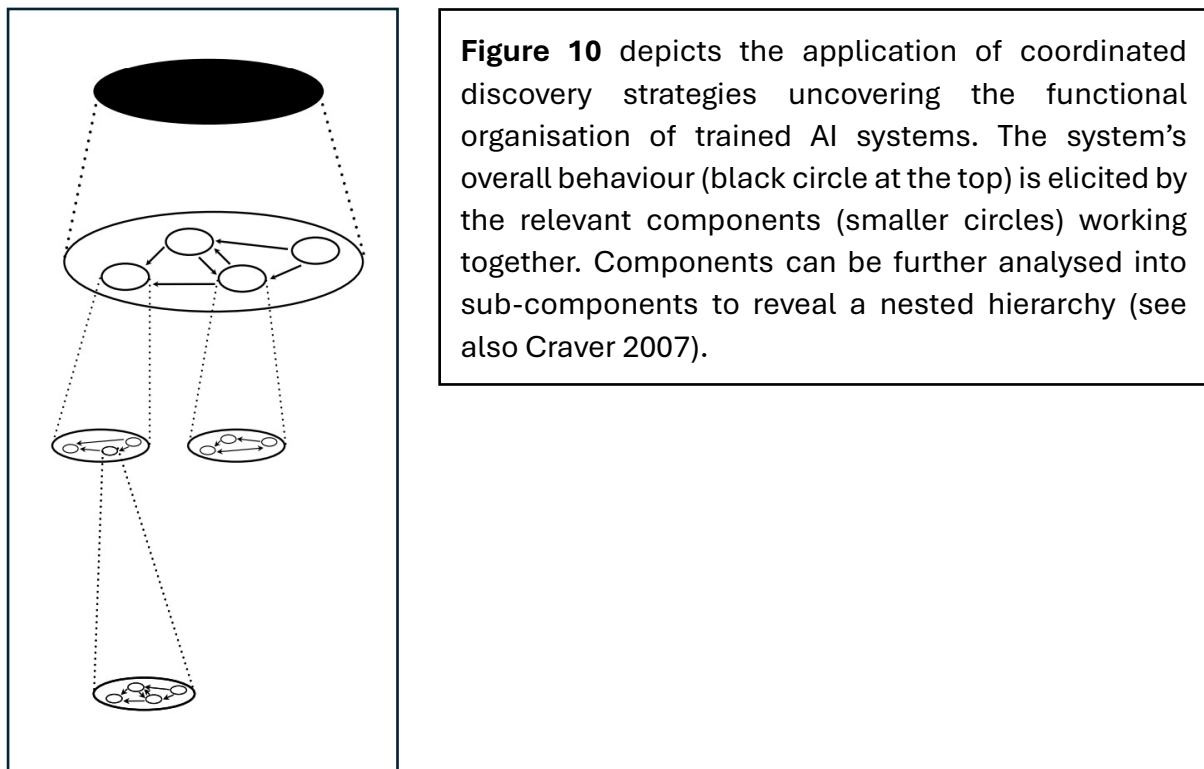
To achieve such characterisations of functional components (EREs) in AI systems, researchers must detect and describe the robust patterns that underlie or maintain the system's behaviour (Wimsatt, 1994). In other words, they must engage in a *pattern recognition practice* (Haugeland, 1998; Kästner & Haueis, 2021), where patterns are any non-random arrangements within systems (Dennett, 1991) which (in virtue of their orderly character) serve as candidates for recognition. As such, patterns constitute potential EREs when seeking MI for AI systems.

Pattern recognition practices in science are collective endeavours involving the coordinated application of shared skills, tools, and concepts (Brigandt, 2011; Kästner & Haueis, 2021). They consist in a set of *epistemic activities* which conform to epistemic norms shared and continuously refined by the research community throughout scientific inquiry. Epistemic activities include such things as *decomposition* and *localisation* (cf. Bechtel & Richardson, 1993; Brigandt, 2011; Kästner & Haueis, 2021) or *recomposition* (cf. Bechtel & Abrahamsen, 2005). *Decomposition* involves breaking a phenomenon down into a set of constitutive sub-functions and is often applied recursively to produce multi-level functional characterisations (Craver, 2007). *Localisation* is the assignment of a sub-function to a particular part of the system which is hypothesised to implement it. *Recomposition* involves building a system back up in a functionally informed manner. To carry out an epistemic activity, researchers will typically engage in a variety of *epistemic operations*. These are the atomic units of scientific inquiry; they consist of concrete actions that track, measure, or manipulate the components of the target system, or generate, transform, and visualise data. In practice, epistemic activities are typically applied in an iterative manner to create a growing store of collective knowledge and fine-

tune hypotheses about which system components are functionally relevant, how they are organised, and how they interact to produce characteristic phenomena.

As such, we argue, pattern recognition practices are ideally suited to produce interpretable, functional characterisations of *how AI systems work as a whole*. They deliver *mechanistic explanations* (Bechtel & Abrahamsen, 2005; Craver, 2001; Kästner & Haueis, 2021) of how the functionally relevant components of a system interact to produce a phenomenon of interest. In the best case, these explanations are comprehensive, providing a detailed, multi-level account of how a system works (see Fig. 10). This holds for artificial systems just as for biological ones.

FIGURE 10: MULTI-LEVEL MECHANISTIC EXPLANATION



3.3.2 The case of Distill

Before turning to a case study, we note that MI has gained significant traction recently, with much work focused on transformer-based LLMs (Bricken et al., 2023; Li et al., 2023; Manning et al., 2020; Meng et al., 2023). One notable approach here is *dictionary learning*, in which sparse autoencoders are used to discover features which are represented across numerous neurons (Bricken et al., 2023; Rajamanoharan et al., 2024). Another particularly promising approach (not specific to transformers) involves

causal abstractions, which enable hypotheses about the functional role of subgraphs of neural networks to be evaluated rigorously through systematic manipulations (Geiger et al., 2022, 2023). Yet another strategy, explicitly inspired by cognitive science, is *representation engineering*, which attempts to locate high-level concepts and processes within trained models (Zou et al., 2023).⁷⁶ Further avenues of active investigation within MI include studying the global representational properties of ANNs (Elhage et al., 2022), how mechanistic structure evolves throughout the learning trajectory (Nanda et al., 2023), and the potential automation of numerous aspects of the MI research process (Conmy et al., 2023; Hernandez et al., 2022; Marks et al., 2024). For useful surveys and discussion of the strengths and weaknesses of some of these methods see R  z (2023) and R  uker and colleagues (2023).⁷⁷ At this point, we wish to reiterate that while certain methods are particularly well-suited for MI, what distinguishes MI from traditional XAI is not the methods it uses, but way it applies methods in coordination with scientific reasoning in pattern recognition practices that are systematic, coordinated, and iterative. As such, for brevity and clarity of exposition of this central point, we limit our philosophical analysis in this paper to the case of Distill.

In their work on the image classification ANN *InceptionV1* (Cammarata et al., 2020; Olah et al., 2018), Chris Olah and colleagues explicitly endorse applying life science research strategies to ANNs, suggesting that “neural networks are an object of empirical investigation, perhaps similar to an organism in biology” (Olah et al., 2020). Hence, they set out to characterise the functional structure of InceptionV1 using coordinated discovery strategies as described above.

For reasons of space, we cannot describe the InceptionV1 architecture in detail here; but a brief introduction to the structure of convolutional ANNs (CNNs) will aid understanding of what follows. As this family of models was designed to process image data, their architecture reflects an attempt to match the compositional and hierarchical properties

⁷⁶ Note that Zou and colleagues frame their approach as an *alternative* to MI, which they construe narrowly as the decomposition of ANNs into circuits. However, on our broader construal of MI as a pattern recognition practice, their approach clearly fits within it.

⁷⁷ The discussion of robust concept detection in R  z (2023, p. 5) is particularly pertinent to our discussion. However, we do not think the functional structure of trained AI systems necessarily needs to be explained in terms of concepts (see Boge, 2023 for discussion of whether ANNs learn concepts).

of natural images (Chollet, 2021). The convolutional layers that give CNNs their name consist of filters which cover a small portion of the input image (represented as an array of raw pixel values). These filters slide over the image, computing activations (dot products of input and filter plus bias) at each spatial position. By analogy with biological neural networks (and mathematical equivalence with other ANN architectures), filters can be thought of as the receptive fields of neurons, with one neuron corresponding to each spatial position the filter covers.⁷⁸ As filters consist of trainable parameters, neurons can learn to respond to different features in the image, producing maps reflecting the presence or absence of features (e.g., edges) to feed to downstream layers. Through composition of simple features, neurons in these downstream layers can thus become sensitive to complex features like shapes, objects, and people. Thus, the very structure of CNNs (their depth, connectivity, and layer design) are adapted to the hierarchical compositionality that characterises images.

Olah and colleagues (Cammarata et al., 2021; Olah, Cammarata, Schubert, et al., 2020) discover and analyse a *curve detector circuit* within InceptionV1. Circuits are sub-graphs of neural networks which, crucially, are not specified as distinct parts of the ANN's architecture. Instead, circuits are part of the model's learned structure. They are functional units which neurons self-organise into during the training process. As such, the curve detector circuit, consisting of a group of neurons spanning five early layers and encompassing around 50 thousand parameters, is a compelling example of an ERE that can only be rendered accessible through a pattern recognition process seeking MI. To uncover this circuit, Olah and colleagues utilise various epistemic activities including decomposition and localisation. They begin by using feature visualisation, a technique in which an image is synthesised to maximise the activation of a given ANN component (Olah et al., 2017).⁷⁹ The Distill team uses the term *features* to mean human interpretable

⁷⁸ In biological brains, each neuron's weights have to be learned separately, even if they are learning to detect the same visual feature (e.g., edges). In ANNs, exploiting the assumption that features may appear anywhere in any image, weights can be shared among neurons for computational efficiency.

⁷⁹ For reasons of space and to retain focus on the pattern recognition practice, we will not go into formal details but instead paint the general picture in relatively broad strokes. For technical details see the referenced work.

concepts that components in ANNs become functionally specialised for.⁸⁰ They taxonomise the neurons in the first 5 convolutional layers of InceptionV1 into layer-wise “families” (i.e., functional groups) based on which features they are sensitive to (e.g., Gabor filters, colour contrasts, lines, curves). Using feature visualisations to represent neurons and families in diagrams of the ANN makes reasoning about their functional relationship significantly more tractable. Notice that this epistemic activity is composed of iterating multiple epistemic operations: producing visualisations for each neuron, applying labels to them, and grouping the neurons into families.

Taxonomising neurons serves as a useful starting point, but a mechanistic explanation of InceptionV1 also requires characterising the *interactions* between neurons. To achieve this, having documented the presence of curve detector neurons in InceptionV1’s third layer, the investigators followed their connections to see how upstream neurons (those closer to the input layer) were contributing to their capabilities. This required the application of additional tools, such as methods to visualise the connection weights between neurons (Voss et al., 2021). Iterating this strategy all the way back to the input layer allowed the researchers to develop a holistic view of the circuit. In so doing, Olah and colleagues uncovered further meso-scale EREs such as *circuit motifs*, which are recurring patterns across multiple neurons. For example, many neurons are “rotationally equivariant”, detecting features (e.g., curves) which are identical apart from their spatial orientation (Olah, Cammarata, Voss, et al., 2020). Ultimately, the Distill team concisely describes the mechanistic structure of the curve circuit as follows: “Gabor filters turn into proto-lines which build lines and early curves. Finally, lines and early curves are composed into curves” (Cammarata et al., 2021). Through coordinated and systematic application of epistemic activities, the full curve circuit becomes recognisable as an ERE distinct from the rest of the network within which it is embedded.

To ensure that discovered patterns are robust, pattern recognition must adhere to epistemic norms such as integrating multiple sources of evidence and ruling out

⁸⁰ The authors conceive of features as corresponding to *directions in neuronal activation space*. In theory, these directions can be defined across the vector space determined by the activity of arbitrary groups of neurons.

alternative explanations.⁸¹ Otherwise, overreliance on visualisation techniques which involve researcher degrees of freedom risks falling foul of common scientific blunders like confirmation bias (Freiesleben, 2024; Gelman & Loken, 2019; Pu & Kay, 2018). In line with this, the Distill researchers used multiple techniques to verify their hypotheses. For example, they observed the responses of curve detector neurons to natural images to confirm they played their hypothesised roles with real stimuli. However, this raises the possibility of a rival hypothesis; could the curve detectors actually specialise in detecting a finite set of specific curved objects, rather than the more general sub-function of *curve detection*? To exclude this rival hypothesis, Olah and colleagues systematically tested the neurons with synthetic stimuli, determining that the curve detection neurons in layer 3b of InceptionV1 are invariant to both fill and colour (i.e., they really are curve detectors). In addition to applying decomposition and localisation to the trained network, Olah and colleagues carried out a (computational) recomposition of the curve detection circuits they had investigated (Cammarata et al., 2021). To do this, one of the authors, leveraging the insights gleaned from the discovery process, designed a curve detection algorithm with the same structure as InceptionV1's by hand. That is to say, they manually set the values of the weights between the neurons. Having done so, the behaviour of the hand-designed network could be compared to that of InceptionV1 by exposing them to identical stimuli and observing their responses. A highly similar response profile was demonstrated, suggesting that Olah and colleagues really had recomposed the curve detection circuit. Indeed, we think this recomposition constitutes a sufficiently *severe test* to be confident in Olah and colleagues' hypothesis of the functional organisation of the curve circuit (see Lam, 2022 for a similar claim).⁸²

The Distill investigation into InceptionV1 exemplifies all of the properties of a pattern recognition practice, viz., the coordinated application of epistemic operations consisting

⁸¹ For more on robustness see, e.g., Wimsatt (1981).

⁸² A severe test of a hypothesis H, as characterised by Deborah Mayo, is such that "H agrees with the data (H passes the test), [and] also [...] with high probability, H would not have passed the test so well, were H false" (Mayo, 2018, p. 92). We think that the potential for MI to provide severe tests of hypotheses about the internal structure of trained models distinguishes it from the divide-and-conquer strategy (see Geiger et al., 2023 for another approach to MI that arguably implements severe testing).

of a shared set of concepts, tools, and skills.⁸³ Olah and colleagues organised their investigation around core concepts: *features* and *circuits*.⁸⁴ When it comes to tools, various visualisation techniques, synthetic stimuli, and explainability interfaces were developed and employed throughout the discovery process. Finally, implementing a pattern recognition practice on ANNs demands multiple skills. Unlike in a wet lab, manipulation of physical instruments is not required. Rather, a firm grasp of the mathematics underlying ANNs is critical. What *is* shared with discovery processes in the life sciences are the reasoning skills required, i.e., choosing which system properties to track and visualise, abductive reasoning to generate hypotheses, and design of experimental procedures to test those hypotheses.

3.3.3 Benefits of the MI approach

Unlike the divide-and-conquer strategy, the pattern recognition practice applied by Olah and his colleagues followed a coherent, systematic research agenda. The scientists' goal was to uncover functional components within the classifier and to illuminate how they work together in the system's overall functional architecture. To this end, they (i) *systematically* searched for functional units (i.e., EREs) that had not been explicitly coded, (ii) *coordinated* different epistemic activities to help reveal the relationships between different measurements, and (iii) *iteratively* refined hypotheses concerning various EREs, their activities and interactions. Eventually, they stitched together a mechanistic explanation rendering intelligible how the first five layers of InceptionV1 work. With respect to curve detection specifically, the authors note that “although curve detection involves more than 50,000 parameters, those parameters actually implement a simple algorithm that can be read off the weights and described in just a few English sentences” (Cammarata et al., 2021).

⁸³ Note that in some cases the computational methods described above have direct analogues in biology and neuroscience. However, the similarity we wish to evoke holds at the level of *research strategies* not at the level of *individual methods*. For discussion of specific methodological similarities see Ivanova et al. (2021) on how probing methods in MI were derived from multivariate pattern analysis in neuroscience and see, e.g, Bashivan et al. (2019) for use of synthetic visualisations in a neuroscience context. On how interventionist theories of causality can be applied to test causal hypotheses in ANNs see Geiger et al. (2021).

⁸⁴ The investigation of *circuits* as a relevant unit of analysis has been adopted by further researchers (Marks et al., 2024; Wang et al., 2022), highlighting the cumulative nature of research from the scientific perspective.

The case of Distill highlights that MI research has two distinct advantages. First, it can shed light on EREs (e.g., the curve detection circuit) of opaque AI systems that are inaccessible through individual XAI methods. There are two reasons for this. The first is that, like measurement tools in the life sciences, individual XAI methods are tuned to picking up specific kinds of information. While tools specialised for answering specific questions are clearly useful, they are also crucially limited. For instance, EEG measures electrical potentials, but it cannot tell us about brain structure. The second reason is that while a lot of information may be gathered from applying different XAI methods, that information will often stand unconnected. By contrast, successful pattern recognition (yielding ERE identification) crucially depends on the coordinated application of many relevant instruments and the *integration* of multiple sources of evidence in conjunction with careful *scientific reasoning* (cf., Kästner, 2018), potentially involving severe testing to rule out alternative hypotheses (see fn 15). By identifying previously unknown EREs, and confirming they have been characterised accurately (e.g., through recomposition), researchers pursuing MI research will often be rewarded with deeper explanations revealing a wider set of EREs and uncovering the functional architecture of AI systems (see Fig. 10). We think this is precisely what is needed to render AI systems holistically explainable.

This leads us to the second advantage of MI research. Characterising a system's functional structure in terms of how relevant components work together enables domain experts to control (e.g., through targeted interventions) and predict system behaviour even for novel situations (Geiger et al., 2023; Meng et al., 2023; Zou et al., 2023). As such, it helps prevent unexpected failures and thus contributes to satisfying important societal desiderata. We already briefly outlined how MI contributes to system reliability and safety in high-stake domains (in Chapter 3.2). While image recognition may not be such a domain *per se*, the case of Distill serves as proof of principle that a fine-grained characterisation of the functional organisation of complex ML-based systems is possible. Besides, ML-based image recognition is likely to be utilised by AI systems employed in high-stakes domains such as traffic control, law, and surveillance. In addition to ensuring safety, the kind of understanding MI elicits is also crucial to support trust. While the relationship between interpretability, trust, and trustworthiness is

complicated (Markus et al., 2021), it is widely agreed that if AI systems are going to be employed to make socially and morally consequential decisions, they need to be *trustworthy* (Kästner et al., 2021). The trust that laypeople have in technological systems is often founded upon their belief that *someone, somewhere* understands the system deeply (Sloman & Rabb, 2016). For sufficiently complex AI systems, we believe, such deep understanding depends upon uncovering EREs that are not attainable through the divide-and-conquer strategy. By contrast, MI has the potential to produce the kind of expert understanding that supports trust.

Another desideratum for which MI is likely to be important is scientific understanding. ANNs have already had a large impact on many scientific fields (Boge, 2022; Bouatta et al., 2021; Cichy & Kaiser, 2019). They achieved unrivalled predictive success for complex phenomena as diverse as neural activity, protein folding, and particle physics. However, philosophers have expressed concern that ANN models, while undoubtedly impressive, do not confer scientific understanding (Chirimuuta, 2021). This limitation is usually traced to the inscrutable learned structure of these models. However, this structure is *precisely* what MI seeks to characterise – and once it is captured successfully, formerly opaque ML-based systems may be re-engineered as transparent or interpretable ones. Therefore, we expect MI to play an increasingly important role for scientific research in the coming years (see also, Chapter 5).

Though our discussion in this section has focused on the potential of MI, we want to be clear that we are *not* calling to eliminate traditional XAI. Rather, we take both approaches to be *complementary*. MI research requires extensive technical knowledge, scientific skill, financial resources, and time. Just as in life science research, it may take several iterations before a relatively stable mechanistic explanation for a phenomenon is uncovered (cf. Craver & Darden, 2013, Kästner & Haueis, 2021). Thus, seeking MI for AI systems will usually incur much greater investment (both financial and labour) than developing XAI methods and applying them in well-defined contexts. These higher costs might only be worth paying in high-stakes scenarios. Likewise, understanding the overall functional organisation of a system might simply not be relevant in some cases (Durán & Jongsma, 2021). Besides, we think that both MI research and applying traditional XAI methods in a divide-and-conquer fashion will often be *mutually supportive*. Researchers

seeking MI can employ specific XAI methods as tools to support some of their coordinated epistemic activities. Likewise, MI research may help refine specific XAI methods, e.g. by revealing new EREs, or improve the mapping between different XAI methods and contexts for the divide-and-conquer strategy. Thus, to what extent MI research and traditional XAI research will be required or useful will depend on the questions at hand, the stakeholders involved, and the specific desiderata at play. Still, it is important that MI research and its unique potential is not overlooked – neither by practitioners nor theorists of XAI.

3.4 Six Worries ... and Responses

Before closing, we shall briefly outline and respond to some possible objections to the MI approach. Though the list below is not exhaustive, we take it to capture the most common worries our proposal will face. We intentionally keep this brief, as a full discussion would require a distinct project.

First, some critics might worry that the chances of successful MI research are too low. For instance, because (#1) the *complexity of large models* (due to their number of parameters and non-linearities) makes explaining them mechanistically hopeless (e.g., Cearn et al., 2019). Consider GPT3, which contains 175 billion parameters (Brown et al., 2020). Clearly, manually investigating every parameter of such large models one-by-one is infeasible. However, this is not required for MI. Instead, researchers are seeking patterns within the complex system to characterise its function. These patterns may not only involve many parameters, they can also recur multiple times within the system, dramatically reducing effective complexity (recall the ‘neuron families’ and ‘circuit motifs’ discussed in Chapter 3.3.2). There is also the potential for some aspects of MI research to be automated, further ameliorating the problem (Conmy et al., 2023; Hernandez et al., 2022). Besides, models need not be explained exhaustively “all the way down” to parameter-level to accrue important benefits. The diversity of societal desiderata for which mechanistic insights may be relevant precludes specifying general criteria for sufficient mechanistic understanding, but in some cases, relatively abstract descriptions may be sufficient (e.g., Tenney et al., 2019; see also Lindsay & Bau, 2023 on

evaluating understanding of neural systems; and Craver & Kaplan, 2020 on sufficiency conditions for mechanistic explanation more generally). This is well in line with the demands formulated in the EU AI Act (see Chapter 3.2) as well as considerations about explainability and understanding being relative to different stakeholder needs (Langer, Oster, et al., 2021).

Others may worry that (#2) we might be *lacking the right concepts* to accurately characterise an AI system's functional organisation, either in terms of the complex features it exploits (Boge, 2022)⁸⁵ or in terms of how it represents and processes those features. We think that even if this is true for now, it is not an argument against seeking MI in principle. It can make implementing a pattern recognition practice *harder*, but forming new concepts is part of the ordinary business of scientific discovery (cf. Craver & Darden, 2013, chapter 5). Besides, the entire history of the life sciences speaks to humans' ability to characterise unfamiliar and complex domains in various useful ways (Bechtel & Richardson, 1993; S. D. Mitchell, 2002). Indeed, Olah and colleagues borrow and fruitfully apply the concept of *network motifs* from systems biology during their curve circuits investigation (Alon, 2006; Olah, Cammarata, Voss, et al., 2020). Thus, we see little reason (theoretical or empirical) to think that trained AI models contain structure so alien that human beings are *fundamentally* unable to grasp it.⁸⁶ Where concepts are lacking, they can be invented (see Schubert et al., 2021 for an example of this in Distill's work). Science frequently operates by making domains interpretable through visualisation, simplification, idealisation, and abstraction (Fleisher, 2022; Levy & Bechtel, 2013). To be sure, the conceptual challenge here should not be taken lightly, but we think the state of evidence suggests that MI is at least worth attempting.

Besides, a critic might protest that (#3) the analogy to life sciences research is flawed because *we do not actually understand how living systems work either*. Indeed, many

⁸⁵ Boge's concern here is with *w-opacity*, that is, opacity with respect to whatever information hidden in the data the trained system has learned to exploit. W-opacity is thus narrower than the functional characterisation targeted by MI, which *also* involves specifying *how* the subparts of a system encode and process that learned information to generate behaviour. In other words: w-opacity is part of the problem MI aims to overcome.

⁸⁶ We acknowledge that this is a complicated issue and much turns on how one construes "grasp" in this sense. However, discussing this in detail would make for another paper. For discussions on the concept of grasping see, e.g., Baumberger et al. (2016) and Janvid (2018).

neuroscientists admit that we do not yet have a comprehensive mechanistic understanding of how the brain works despite decades of effort (Buzsáki, 2020; Pessoa, 2023). Yet, we see reasons to be *more* optimistic about understanding AI systems through MI than explaining the brain: (i) since AI systems are designed and trained by us, we have exhaustive knowledge about their architectures, learning algorithms, objective functions, training data, and learned connection weights, (ii) we can perform any conceivable experiment with arbitrary precision and complete access to the causal consequences of our interventions, and (iii) we can choose to develop AI systems which are tailored to be amenable to scientific investigation (e.g., by enforcing linearity wherever it does not significantly hurt model performance). While none of these advantages make applying MI to complex AI systems easy, we think that in combination they warrant cautious optimism.

Similarly, critics may worry that (#4) the analogy to life science research is flawed because *ANNs defy decomposition into components with different functional roles*. We grant that, given the distributed nature of ANNs, one should not assume subfunctions will map neatly onto pre-individuated structures within an ANN (neurons, layers, etc.). However, as discussed in Chapter 3.3.1, this is not a pre-requisite for MI to prove valuable. Identifying components relevant for a specific function may involve characterising more exotic and distributed structures (Bricken et al., 2023). This may not be straightforward, but that is another worry (akin to #1 and #2). The exact criteria a system must satisfy to afford effective mechanistic discovery processes is a complex topic we cannot address in this paper (for classic and recent discussions see Simon, 2008; Woodward, 2013; Zednik, 2015). Broadly speaking though, we think extant evidence (including the Distill case study) strongly suggests that ANNs and other modern AI systems tend to be decomposable *enough* for MI to provide value. Further, as Zednik (2015) points out, the ongoing production of more powerful methods and strategies for conducting mechanistic discovery processes continually expands the range of systems that are usefully decomposable. Overall, we consider decomposability a matter of degree (cf. Bechtel & Richardson 1993). And precisely how (and how easily) decomposable specific ANNs turn out to be is an empirical question (that only seriously attempting MI can possibly answer).

Even if MI research is successful, some might worry that the benefits we highlighted above (see Chapters 3.2 and 3.3.3) will not be worth the trouble. For instance, those sympathetic to our approach in principle may argue that (#5) *it will be impossible for researchers to seek MI for large AI systems as quickly as they are developed*. It would just be too resource intensive. While we acknowledge MI research requires significant resources (see Chapter 3.3.3), we think there are good reasons to pursue it anyway. First, it might simply be worth the investment if otherwise we cannot safely deploy AI in high-stakes scenarios (see Chapter 3.2). Second, because mechanistic insights may be transferable from one system to another, the investment might pay off more quickly than we think. For instance, insights about the functional organisation of one language model may well transfer to other language models (*mutatis mutandis* for other domains). Theoretical reasons for this optimism are provided by Cao and Yamins' (2021) *contravariance principle*, which implies that functional convergence should be expected in systems that can perform complex tasks efficiently (see also Bansal et al., 2021; Schrimpf et al., 2020). Besides, researchers are increasingly converging on architectural choices for AI systems (Bommasani et al., 2022). Since model architecture constrains both the kind of functional structure that can emerge and the kind of epistemic operations researchers can apply, this is positive news for MI. Moreover, the resource intensive nature of pattern recognition practices can be viewed as a feature, rather than a bug. There is still understandable resistance to large AI systems being deployed in high-stakes scenarios (Rudin, 2019). In light of these concerns, it may be reasonable to propose some degree of mechanistic explainability as a pre-requisite for adopting technological innovations that could prove consequential and difficult to reverse (Stirling, 2007). A time-consuming requirement could give slow political processes time to react and inject democratic influence into decisions that impact all of society (Bender et al., 2021; Floridi et al., 2018).

Other critics may worry that (#6) *MI will not yield accurate predictions of AI systems' behaviours in social contexts* – though this is what really matters when we think about deploying AI in modern society (Bender et al., 2021). We acknowledge MI will not be sufficient to precisely predict all effects of deploying models in specific real-world contexts with idiosyncratic data distributions (Quiñonero-Candela, 2009; such

prediction is generally intractable, see Shalizi, 2006). However, given MI research seeks to reveal how AI systems work as a whole, it yields deeper and more generalisable insights into how a system may behave in novel (social) contexts than applying the divide-and-conquer strategy. Further, the ability to intervene on trained systems, which MI purports to deliver, may prove important for controlling AI systems after they are deployed. Thus, even if MI is costly and the benefits are not all-encompassing, it may still be superior to its alternatives.

Before concluding, we wish to stress that our aim in this paper is to characterise MI as a distinctive approach to XAI, not to present it as a panacea. As such, though we reject the notion that any of these objections is fatal to the MI research program, we acknowledge that they are legitimate considerations which must be factored into a comprehensive assessment of its pursuit worthiness in specific cases. In a nutshell, MI research can offer answers to research questions which are in principle unavailable to those pursuing a divide-and-conquer strategy (see Chapter 3.2). As such, it is important for practitioners to pursue MI to gain explainability where other strategies are unproductive. And it is also crucial for theorists to accommodate MI in their in their portrayals of the XAI research landscape (see Chapter 3.3.3). For only by having the full range of strategies for explaining AI systems in view can we make appropriate choices about which research to pursue to ensure societal desiderata are satisfied.

3.5 Conclusion

Contemporary XAI research is seeking explainability for opaque AI systems in a divide-and-conquer fashion, viz. by employing specific XAI methods in various contexts. While this strategy has its merits, it fails to illuminate how trained AI systems work as a whole. Yet precisely this kind of holistic understanding is needed to satisfy important desiderata placed on AI systems by society. We argued that mechanistic interpretability research, though it is resource intensive, can serve as a remedy: by uncovering a wider range of EREs than any individual XAI method, MI contributes to a deeper and more holistic understanding of complex AI systems' functioning. It aims to uncover the functional organisation of trained AI systems, including both the complex features they exploit and

how they represent and process those features. These insights can be utilised to predict and control system behaviour beyond pre-defined contexts, which is crucial to satisfy societal desiderata. Appreciating these unique benefits of MI research will help scholars to make adequate choices about what research to pursue to ensure we can safely and reliably deploy ML-based systems in modern society.

3.6 References

- Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781420011432>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. <https://doi.org/10.48550/arXiv.1606.06565>
- Artificial Intelligence Act (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf
- Baetu, T. M. (2011). Mechanism schemas and the relationship between biological theories. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences* (pp. 407–424). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199574131.003.0019>
- Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting Model Stitching to Compare Neural Representations. *Advances in Neural Information Processing Systems*, 34, 225–236. <https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html>
- Baraniuk, R., Donoho, D., & Gavish, M. (2020). The science of deep learning. *Proceedings of the National Academy of Sciences*, 117(48), 30029–30032. <https://doi.org/10.1073/pnas.2020596117>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science (New York, N.Y.)*, 364(6439), eaav9436. <https://doi.org/10.1126/science.aav9436>
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. 3319–3327. <https://doi.org/10.1109/CVPR.2017.354>
- Baumberger, C., Beisbart, C., & Brun, G. (2016). What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science. In *Explaining Understanding*. Routledge.

- Bechtel, W. (2009). Looking Down, Around, and Up: Mechanistic Explanation in Psychology. *Philosophical Psychology*, 22. <https://doi.org/10.1080/09515080903238948>
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–441. <https://doi.org/10.1016/j.shpsc.2005.03.010>
- Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research* (pp. xiv, 286). Princeton University Press.
- Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4. <https://www.frontiersin.org/articles/10.3389/fdata.2021.688969>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1), 43–75. <https://doi.org/10.1007/s11023-021-09569-4>
- Boge, F. J. (2023). Functional Concept Proxies and the Actually Smart Hans Problem: What's Special About Deep Neural Networks in Science. *Synthese*, 203(1), 16. <https://doi.org/10.1007/s11229-023-04440-8>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. <https://doi.org/10.48550/arXiv.2108.07258>
- Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section D, Structural Biology*, 77(Pt 8), 982–991. <https://doi.org/10.1107/S2059798321007531>
- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.
- Brigandt, I. (2011). Explanation in Biology: Reduction, Pluralism, and Explanatory Aims. *Science & Education*, 22(1), 69–91. <https://doi.org/10.1007/s11191-011-9350-7>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>

- Burnston, D. C. (2021). Getting over Atomism: Functional Decomposition in Complex Neural Systems. *The British Journal for the Philosophy of Science*, 72(3), 743–772. <https://doi.org/10.1093/bjps/axz039>
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512. <https://doi.org/10.1177/2053951715622512>
- Buzsáki, G. (2020). The Brain–Cognitive Behavior Problem: A Retrospective. *eNeuro*, 7(4). <https://doi.org/10.1523/ENEURO.0069-20.2020>
- Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, 5(3), 10.23915/distill.00024. <https://doi.org/10.23915/distill.00024>
- Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve Circuits. *Distill*, 6(1), e00024.006. <https://doi.org/10.23915/distill.00024.006>
- Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. *arXiv:2104.01489 [Cs, q-Bio]*. <http://arxiv.org/abs/2104.01489>
- Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, 9(1), Article 1. <https://doi.org/10.1038/s41398-019-0607-2>
- Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199(1), 767–790. <https://doi.org/10.1007/s11229-020-02713-0>
- Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Simon and Schuster.
- Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4), 305–317. <https://doi.org/10.1016/j.tics.2019.01.009>
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). *Towards Automated Circuit Discovery for Mechanistic Interpretability* (arXiv:2304.14997). arXiv. <https://doi.org/10.48550/arXiv.2304.14997>
- Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, 68(1), 53–74.
- Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.
- Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press. <https://press.uchicago.edu/ucp/books/book/chicago/l/bo16123713.html>
- Craver, C. F., & Kaplan, D. M. (2020). Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. *The British Journal for the Philosophy of Science*, 71(1), 287–319. <https://doi.org/10.1093/bjps/axy015>
- Crook, B., & Kästner, L. (Forthcoming). Don’t Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Springer.
- Dennett, D. C. (1991). Real Patterns. *Journal of Philosophy*, 88(1), 27–51. <https://doi.org/10.2307/2027085>

- Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics-2020-106820. <https://doi.org/10.1136/medethics-2020-106820>
- Eden, A. H. (2007). Three Paradigms of Computer Science. *Minds and Machines*, 17(2), 135–167. <https://doi.org/10.1007/s11023-007-9060-8>
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. In *arXiv e-prints*. <https://doi.org/10.48550/arXiv.2209.10652>
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., ... Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.
- Fawzi, A., Fawzi, H., & Fawzi, O. (2018). Adversarial vulnerability for any classifier. *Advances in Neural Information Processing Systems*, 31. <https://papers.nips.cc/paper/2018/hash/851ddf5058cf22df63d3344ad89919cf-Abstract.html>
- Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, 19(4), 534–560. <https://doi.org/10.1017/epi.2022.39>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Freiesleben, T. (2024). *Artificial Neural Nets and the Representation of Human Concepts* (arXiv:2312.05337). arXiv. <https://doi.org/10.48550/arXiv.2312.05337>
- Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, 34, 9574–9586. <https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html>
- Geiger, A., Potts, C., & Icard, T. (2023). *Causal Abstraction for Faithful Model Interpretation* (arXiv:2301.04709). arXiv. <https://doi.org/10.48550/arXiv.2301.04709>
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing Causal Structure for Interpretable Neural Networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338. <https://proceedings.mlr.press/v162/geiger22a.html>
- Gelman, A., & Loken, E. (2019). *The garden of forking paths: Why multiple comparisons can be a problem , even when there is no “ fishing expedition ” or “ p-hacking ” and the research hypothesis was posited ahead of time **. <https://www.semanticscholar.org/paper/The-garden-of-forking-paths-%3A-Why-multiple-can-be-a-Gelman-Loken/b63e25900013605c16f4ad74c636cfbd8e9a3e8e>
- Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3), 10.23915/distill.00030. <https://doi.org/10.23915/distill.00030>

- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 93:1-93:42. <https://doi.org/10.1145/3236009>
- Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. <https://doi.org/10.1145/3593013.3594067>
- Haugeland, J. (1998). *Having Thought: Essays in the Metaphysics of Mind*. Harvard University Press.
- Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2018). Deep Anomaly Detection with Outlier Exposure. *ArXiv*. <https://www.semanticscholar.org/paper/Deep-Anomaly-Detection-with-Outlier-Exposure-Hendrycks-Mazeika/6cf1d69e447e9687dbd2d92572f44bdddabd8192>
- Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks* (arXiv:2306.12001). arXiv. <https://doi.org/10.48550/arXiv.2306.12001>
- Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). *Natural Language Descriptions of Deep Visual Features* (arXiv:2201.11114). arXiv. <https://doi.org/10.48550/arXiv.2201.11114>
- Howick, J., Glasziou, P., & Aronson, J. K. (2010). Evidence-based mechanistic reasoning. *Journal of the Royal Society of Medicine*, 103(11), 433–441. <https://doi.org/10.1258/jrsm.2010.100146>
- Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, 471, 61–71. <https://doi.org/10.1016/j.canlet.2019.12.007>
- Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626. <https://doi.org/10.1007/s11229-008-9435-2>
- Ivanova, A. A., Hewitt, J., & Zaslavsky, N. (2021). *Probing artificial neural networks: Insights from neuroscience* (arXiv:2104.08197). arXiv. <https://doi.org/10.48550/arXiv.2104.08197>
- Janvid, M. (2018). Getting a Grasp of the Grasping Involved in Understanding. *Acta Analytica*, 33(3), 371–383. <https://doi.org/10.1007/s12136-018-0348-5>
- Jugovac, M., & Jannach, D. (2017). Interacting with Recommenders—Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems*, 7(3), 10:1-10:46. <https://doi.org/10.1145/3001837>
- Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2022). Towards Causal Algorithmic Recourse. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 139–166). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_8
- Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, 68, 68–79. <https://doi.org/10.1016/j.shpsa.2018.01.011>

- Kästner, L., & Haueis, P. (2021). Discovering Patterns: On the Norms of Mechanistic Inquiry. *Erkenntnis*, 86(6), 1635–1660. <https://doi.org/10.1007/s10670-019-00174-7>
- Kästner, L., Langer, M., Lazar, V., Schomacker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 169–175. <https://doi.org/10.1109/REW53955.2021.00031>
- Keil, F. (2019). *How Do Partial Understandings Work?* (pp. 191–208). <https://doi.org/10.1093/oso/9780190860974.003.0010>
- Keogh, E., & Mueen, A. (2017). Curse of Dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_192
- Kirsch, A. (2017). *Explain to whom? Putting the User in the Center of Explainable AI*. CEx@AI*IA. <https://www.semanticscholar.org/paper/Explain-to-whom-Putting-the-User-in-the-Center-of-Kirsch/da43f222f009a587381a2aa8e145f5a45c26db9f>
- Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368. <https://doi.org/10.1109/RE.2019.00046>
- Lam, N. (2022). Explanations in AI as Claims of Tacit Knowledge. *Minds and Machines*, 32(1), 135–158. <https://doi.org/10.1007/s11023-021-09588-1>
- Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, 29(2), 154–169. <https://doi.org/10.1111/ijsa.12325>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sasing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), Article 1. <https://doi.org/10.1038/s41467-019-08987-4>
- Levy, A., & Bechtel, W. (2013). Abstraction and the Organization of Mechanisms. *Philosophy of Science*, 80(2), 241–261.
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task* (arXiv:2210.13382). arXiv. <https://doi.org/10.48550/arXiv.2210.13382>
- Lindsay, G. W., & Bau, D. (2023). Testing methods of neural systems understanding. *Cognitive Systems Research*, 82, 101156. <https://doi.org/10.1016/j.cogsys.2023.101156>
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054. <https://doi.org/10.1073/pnas.1907367117>

- Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). *Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models* (arXiv:2403.19647). arXiv. <https://doi.org/10.48550/arXiv.2403.19647>
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781107286184>
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2023). *Locating and Editing Factual Associations in GPT* (arXiv:2202.05262). arXiv. <https://doi.org/10.48550/arXiv.2202.05262>
- Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation Studies in Artificial Neural Networks* (arXiv:1901.08644). arXiv. <https://doi.org/10.48550/arXiv.1901.08644>
- Millière, R., & Buckner, C. (2024). *A Philosophical Introduction to Language Models -- Part I: Continuity With Classic Debates* (arXiv:2401.03910). arXiv. <https://doi.org/10.48550/arXiv.2401.03910>
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120>
- Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy*, 17(1), 55–70. <https://doi.org/10.1023/A:1012990030867>
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. <https://doi.org/10.1145/3287560.3287574>
- Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). *Progress measures for grokking via mechanistic interpretability* (arXiv:2301.05217). arXiv. <https://doi.org/10.48550/arXiv.2301.05217>
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126. <https://doi.org/10.1145/360018.360022>
- Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3–5), 393–444. <https://doi.org/10.1007/s11257-017-9195-0>
- Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(1), 13. <https://doi.org/10.1007/s10676-022-09632-3>
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3), e00024.001. <https://doi.org/10.23915/distill.00024.001>

- Olah, C., Cammarata, N., Voss, C., Schubert, L., & Goh, G. (2020). Naturally Occurring Equivariance in Neural Networks. *Distill*, 5(12), e00024.004. <https://doi.org/10.23915/distill.00024.004>
- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, 2(11), 10.23915/distill.00007. <https://doi.org/10.23915/distill.00007>
- Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010. <https://doi.org/10.23915/distill.00010>
- Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>
- Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251), 20220041. <https://doi.org/10.1098/rsta.2022.0041>
- Pessoa, L. (2023). The Entangled Brain. *Journal of Cognitive Neuroscience*, 35(3), 349–360. https://doi.org/10.1162/jocn_a_01908
- Pu, X., & Kay, M. (2018). The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Paper. *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, 37–45. <https://doi.org/10.1109/BELIV.2018.8634103>
- Quiñonero-Candela, J. (Ed.). (2009). *Dataset shift in machine learning*. MIT Press.
- Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). *On the expressive power of deep neural net-works*. <https://arxiv.org/pdf/1606.05336.pdf>
- Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., & Nanda, N. (2024). *Improving Dictionary Learning with Gated Sparse Autoencoders* (arXiv:2404.16014). arXiv. <https://doi.org/10.48550/arXiv.2404.16014>
- Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks* (arXiv:2207.13243). arXiv. <https://doi.org/10.48550/arXiv.2207.13243>
- Răz, T. (2023). Methods for identifying emergent concepts in deep neural networks. *Patterns*, 4(6), 100761. <https://doi.org/10.1016/j.patter.2023.100761>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. 97–101. <https://doi.org/10.18653/v1/N16-3020>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Article 1. <https://doi.org/10.1609/aaai.v32i1.11491>
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. <http://aima.cs.berkeley.edu/>
- Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3), 413–423. <https://doi.org/10.1016/j.neuron.2020.07.040>
- Schubert, L., Voss, C., Cammarata, N., Goh, G., & Olah, C. (2021). High-Low Frequency Detectors. *Distill*, 6(1), e00024.005. <https://doi.org/10.23915/distill.00024.005>
- Shalizi, C. R. (2006). Methods and Techniques of Complex Systems Science: An Overview. In T. S. Deisboeck & J. Y. Kresh (Eds.), *Complex Systems Science in Biomedicine* (pp. 33–114). Springer US. https://doi.org/10.1007/978-0-387-33532-2_2
- Simon, H. A. (2008). *The sciences of the artificial* (3. ed., [Nachdr.]). MIT Press.
- Sloman, S. A., & Rabb, N. (2016). Your Understanding Is My Understanding: Evidence for a Community of Knowledge. *Psychological Science*, 27(11), 1451–1460. <https://doi.org/10.1177/0956797616662271>
- Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. <https://doi.org/10.1145/3351095.3372870>
- Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. <https://doi.org/10.1145/3531146.3534639>
- Steimers, A., & Schneider, M. (2022). Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*, 19(6), 3641. <https://doi.org/10.3390/ijerph19063641>
- Stirling, A. (2007). Risk, precaution and science: Towards a more constructive policy debate. Talking point on the precautionary principle. *EMBO Reports*, 8(4), 309–315. <https://doi.org/10.1038/sj.embor.7400953>
- Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, 44(3), 510–515. <https://doi.org/10.1016/j.shpsa.2012.12.005>
- Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovered the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. <https://doi.org/10.18653/v1/P19-1452>
- Voss, C., Cammarata, N., Goh, G., Petrov, M., Schubert, L., Egan, B., Lim, S. K., & Olah, C. (2021). Visualizing Weights. *Distill*, 6(2), e00024.007. <https://doi.org/10.23915/distill.00024.007>
- Vredenburg, K. (2022). The Right to Explanation*. *Journal of Political Philosophy*, 30(2), 209–229. <https://doi.org/10.1111/jopp.12262>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3063289>

- Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small* (arXiv:2211.00593). arXiv. <https://doi.org/10.48550/arXiv.2211.00593>
- Wei, A., Haghtalab, N., & Steinhardt, J. (2023). *Jailbroken: How Does LLM Safety Training Fail?* (arXiv:2307.02483). arXiv. <https://doi.org/10.48550/arXiv.2307.02483>
- Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination (1981). In L. Soler, E. Trizio, T. Nickles, & W. Wimsatt (Eds.), *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science* (pp. 61–87). Springer Netherlands. https://doi.org/10.1007/978-94-007-2759-5_2
- Wimsatt, W. C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy Supplementary Volume*, 20, 207–274. <https://doi.org/10.1080/00455091.1994.10717400>
- Winning, J., & Bechtel, W. (2018). Rethinking Causality in Biological and Neural Mechanisms: Constraints and Control. *Minds and Machines*, 28(2), 287–310. <https://doi.org/10.1007/s11023-018-9458-5>
- Woodward, J. (2013). Mechanistic Explanation: Its Scope and Limits. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 87, 39–65.
- Woodward, J. (2017). Explanation in Neurobiology: An Interventionist Perspective. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780199685509.003.0004>
- Zednik, C. (2015). Heuristics, Descriptions, and the Scope of Mechanistic Explanation. In P.-A. Braillard & C. Malaterre (Eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences* (pp. 295–318). Springer Netherlands. https://doi.org/10.1007/978-94-017-9822-8_13
- Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288. <https://doi.org/10.1007/s13347-019-00382-7>
- Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5), Article 5. <https://doi.org/10.3390/electronics10050593>
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., ... Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency* (arXiv:2310.01405). arXiv. <https://doi.org/10.48550/arXiv.2310.01405>

Chapter 4:

Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)

This chapter was published as “Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW) (pp. 316-324). IEEE.” Slight changes have been made to fit the dissertation format.

4.0 Abstract

Within the field of Requirements Engineering (RE), the increasing significance of Explainable Artificial Intelligence (XAI) in aligning AI-supported systems with user needs, societal expectations, and regulatory standards has garnered recognition. In general, explainability has emerged as an important nonfunctional requirement that impacts system quality. However, the supposed trade-off between explainability and performance challenges the presumed positive influence of explainability. If meeting the requirement of explainability entails a reduction in system performance, then careful consideration must be given to which of these quality aspects takes precedence and how to compromise between them. In this paper, we critically examine the alleged trade-off. We argue that it is best approached in a nuanced way that incorporates resource availability, domain characteristics, and considerations of risk. By providing a foundation for future research and best practices, this work aims to advance the field of RE for AI.

4.1 Introduction

The field of Requirements Engineering (RE) plays a pivotal role in ensuring that artificial computing systems are aligned with user needs, societal expectations, and regulatory standards. As Artificial Intelligence (AI) systems become increasingly prevalent in modern society, ensuring that these systems align with customer expectations is of

utmost importance. In this regard, Explainable AI (XAI) has emerged as a crucial factor in the pursuit of building transparent and accountable AI systems (Barredo Arrieta et al., 2020; Baum et al., 2022; Langer et al., 2021). In RE, the concept of explainability has received a lot of attention recently, and it is rapidly establishing itself as a vital non-functional requirement (NFR) (Brunotte et al., 2022; Chazette et al., 2021; Chazette & Schneider, 2020; Köhl et al., 2019). As the demand for explainability continues to grow, it is imperative for the RE community to understand this NFR in detail. In this paper, we aim to contribute to this understanding by critically examining the often-proclaimed trade-off between performance and explainability in AI and considering its implications for RE.⁸⁷ The trade-off captures the idea that AI models which achieve better performance are less explainable, potentially leading to misalignment with user needs.

While the trade-off is often taken for granted in the literature (e.g., Barredo Arrieta et al., 2020; Došilović et al., 2018; Gunning et al., 2019), there are also voices that question its existence and worry that unreflective acceptance of it can lead to negative consequences (e.g., Gosiewska et al., 2021; Herzog, 2022; Rudin, 2019). Thus, the question arises as to whether this trade-off actually exists and, if so, how to reconcile it with user expectations. This article aims to explore the performance-explainability trade-off in the context of AI from the perspective of RE. Contrary to polarised viewpoints which either assert the trade-off as a simple fact or deny its existence altogether, we argue for a more nuanced picture that incorporates development constraints, variations in domain complexity, and an accurate appraisal of the potential of XAI. This allows us to clarify the relationship between performance and explainability. Overall, our aim is to provide a foundation for future research and best practices in RE for AI.

We proceed as follows. In Chapter 4.2, we will describe and contextualise the performance-explainability trade-off more fully. First, we introduce the rationale for the trade-off, before turning to arguments presented against it in an influential paper by Cynthia Rudin (2019). Having briefly introduced her argument, Chapter 4.3 is devoted to

⁸⁷ The name given to the trade-off varies; some speak of an accuracy-interpretability trade-off (e.g., Rudin, 2019). We choose the term performance over accuracy for maximal generality. Strictly speaking, accuracy is just one of a number of performance metrics that researchers might choose to evaluate a model. Also, we will only speak of explainability and remain agnostic as to whether there are any differences to interpretability (see Clinciu & Hastie, 2019, Sterz et al., 2021, and Speith, 2022 for discussions).

a critique of it. By examining several aspects of her argument in detail, we show that, while her criticism contains important insights, the view of the relationship between performance and explainability she ultimately presents is incomplete and misleading. We end the article in Chapter 4.4 by presenting our positive vision for how the RE community should think about the trade-off.

4.2 The Performance-Explainability Trade-Off

While the performance-explainability trade-off is often presented as a fact, the reality is more complicated (Bell et al., 2022). In this section, we describe the trade-off in more detail and summarise Rudin's critique of it.

4.2.1 The rationale for the trade-off

In general, large and complex models, such as Deep Neural Networks (DNNs) or ensemble methods, often achieve high scores on performance metrics such as raw predictive accuracy, F1 score, or (negative) log loss. This pattern can be observed in many ML breakthroughs of the past decades, e.g., in computer vision (Farabet et al., 2013; He et al., 2016; Krizhevsky et al., 2012; Lawrence et al., 1997; Lecun et al., 1998), speech recognition (Hinton et al., 2012), natural language processing (Bubeck et al., 2023; Sutskever et al., 2014; Vaswani et al., 2017), and reinforcement learning (Silver et al., 2016; Vinyals et al., 2019). However, these models tend to suffer from a lack of explainability, as it is difficult to understand how they arrive at their decisions (Sullivan, 2022). Since models like DNNs have many parameters with values learned through an automated training procedure, they function as “black boxes”, where their decision-making processes are not easily expressible in human terms (Guidotti et al., 2018; Zednik, 2021). On the other hand, models that are inherently explainable (i.e., ante-hoc explainable models; Speith, 2022a), such as rule-based systems, decision trees or linear classifiers, are simpler and more transparent. Generally speaking, they employ explicit rules or engineered features for their decisions, making it easier for humans to understand the model's reasoning process.

There are several factors which contribute to the explainability of any particular model (Mann et al., 2023). One obvious factor is size, with larger models often being assumed

to be more difficult to understand (Freitas, 2014). Another factor to consider is the type of representations a model uses (Huysmans et al., 2011). Certain representations, such as the nodes of decision trees, tend to be more understandable compared to the distributed representations learned by DNNs (Samek et al., 2019). A final, related factor is the complexity of the interaction between the features of a model. It is easier to understand the contribution of features in a linear classifier, for instance, compared to comprehending how features in a DNN, which may interact non-linearly, influence the model's behaviour (Mittelstadt et al., 2019). Thus, based on the model that one chooses, it seems like one has to trade-off achieving the best possible performance against having a model that is fully explainable (Barredo Arrieta et al., 2020; Došilović et al., 2018; Gunning et al., 2019).

Although the existence of a performance-explainability trade-off in theory (e.g., for artificial data sets) is interesting, the relevant question for RE is whether we must trade these qualities off against one another in real-world contexts. Given this crucial proviso, we can express the claim of the Performance-Explainability Trade-off (PET) as follows:

(PET) *For some real-world tasks, when developing a model to perform (part of) the task, increasing the performance of the model necessarily decreases the explainability of the model (and vice versa).*

4.2.2 Rudin's critique of the trade-off

Rudin argues that, when it comes to high-stakes decisions, we are better off developing inherently explainable AI systems rather than using black-box models and explaining them after they are trained (i.e., post-hoc explainability; Speith, 2022a; Rudin, 2019). She provides several interconnected reasons for adopting this position. Initially, Rudin agrees with proponents of the PET on the premise that black-box models like DNNs are not (ante-hoc) explainable. However, unlike XAI researchers, she takes a dim view of the prospects of explaining these models with post-hoc explainability techniques, claiming that the explanatory information these techniques produce is 1) not faithful to the original models, 2) incomplete, and 3) unhelpful in most contexts. Furthermore, Rudin thinks the performance of simpler models can be improved using techniques like knowledge discovery (Maimon & Rokach, 2005) and feature engineering (Liu & Motoda, 1998).

Crucially, according to Rudin, the trend towards using such opaque models arises from an unwarranted belief in the PET.

Against the PET, Rudin argues that explainable models can be as good as or even outperform black-box models in most real-world applications. While some authors argue that the use of complex models is justified because they find “hidden patterns” in data (Bengio et al., 2013; Ha & Jeong, 2021), Rudin argues that any patterns required for high performance could also be exploited by an ante-hoc explainable model. In the most general statement of her argument, Rudin claims that, for any given task, the set of almost-equally-performant models typically includes at least one simple and explainable model. This so-called *Rashomon Set argument* supports Rudin's overall proposal, which is that we should spend more time researching and engineering inherently explainable models. In Rudin's view, this will not lead to lower performance, and thus avoids the trade-off.

Our aim is not to directly disagree with Rudin's conclusion, but rather to address some issues in the reasoning about the PET she employs to support her claims. In particular, we will critically discuss Rudin's stance on three topics which contribute to a nuanced appraisal of the PET: 1) the potential of post-hoc explainability methods, 2) the limits of iterative feature engineering, and 3) the scope and plausibility of the Rashomon Set argument.

4.3 Contra Rudin's Critique

We will now critically analyse Rudin's argument. To do so, we will examine each of her subclaims individually.

4.3.1 The potential of post-hoc explainability approaches

First, we want to examine Rudin's assessment of the potential of post-hoc explainability approaches. We think that she underestimates this potential in terms of three properties, namely, the approaches' fidelity, completeness, and usefulness.

Fidelity

Let us first discuss Rudin's take on the fidelity of post-hoc explainability approaches. Following her discussion, we assume that fidelity means that an explanation accurately captures the decision processes of a model (Speith, 2022b); i.e., that it reflects how and what the model computes (Molnar, 2022). In her argument, Rudin asserts that explaining an AI model after it has been trained (i.e., post-hoc explainability) must result in explanations that lack fidelity to the original model. Further, she argues that these explanations should not really be called “explanations” at all, since they could respond to completely different features than the original model. According to her, this is not morally acceptable, especially in situations where consequential human values are at stake. For example, a person might be denied medical treatment due to a racial bias in the model, but this could be (un)intentionally obscured by explanations that are not fidelitous to the model.

While Rudin's concerns may hold true for model-agnostic explainability approaches, i.e., approaches that explain a model without regards to its internals or structure, e.g., by mere input/output analysis (Speith, 2022a), it fails to consider the potential of model-specific ones, i.e., approaches that explain a model with at least some recourse to the model's internals or structure (Speith, 2022a) to provide fidelitous explanations. It is true that there is a lack of research that addresses the fidelity of information generated by explainability approaches. This research gap can be attributed, in part, to the absence of formal and technical criteria for assessing fidelity (Adebayo et al., 2018; Amparore et al., 2021; Molnar, 2022). Recently, however, researchers have started to address this issue (Amorim et al., 2023). Adebayo et al. (2018), for instance, proposed a sanity check specifically designed for approaches that generate saliency maps. Such checks can be used to assess whether explainability approaches are sensitive to class-relevant features, thus providing evidence against the worry that they lack fidelity. By applying the check, Adebayo et al. assessed various approaches and identified variations in fidelity. Notably, while some model-specific approaches such as Vanilla Backpropagation (Simonyan et al., 2014) and Grad-CAM (Selvaraju et al., 2020) demonstrated fidelity by passing the check, others like Guided Backpropagation (Springenberg et al., 2015) did not.

In light of these findings, it becomes evident that Rudin's claim requires a more nuanced perspective. By evaluating the fidelity of explanations it is possible to distinguish between explainability approaches that offer accurate representations of decision-making processes and those that do not. Further exploration and evaluation of model-specific explainability approaches are necessary to gain a deeper understanding of their fidelity and potential to provide reliable insights into AI models.

Completeness

Admittedly, Rudin's criticism may not only concern the fidelity of explanations, but also their completeness (see Zhou et al., 2021 for related discussion): that all factors playing into a model's output are cited. If we analyse the completeness of saliency maps, the situation does not look as good as it did with their fidelity (Dombrowski et al., 2019). This is because saliency maps are mathematically capable of responding only to the most important features, and it cannot be ruled out that some comparably less important features also had an impact on the output of a model. However, such features could be crucial, for example, when it comes to assessing possible biases regarding protected attributes (Tschantz, 2022).

Saliency maps, however, are not the only type of post-hoc explainability approach that is potentially fidelitous. In particular, there are several approaches that attempt to exploit the latent information contained in a DNN for explanations, such as feature visualisation (Olah et al., 2017, 2018) and Testing with Concept Activation Vectors (TCAV) (Kim et al., 2018). Feature visualisation, for instance, aims to make the role of neurons or layers in DNNs clear by synthesising images that maximise their activation (Nguyen et al., 2019; Olah et al., 2017). Through systematic use of these techniques, researchers are able to build up fidelitous and complete explanations of the internal structure of opaque models (Cammarata et al., 2020). Furthermore, incomplete explanations that are fidelitous can be sufficient in many cases, such as when the tracked features suffice to expose a model's deficiency. In these cases, incomplete explanations are even preferable to complete explanations, since the latter can overwhelm a recipient and, in some cases, even obscure the malfunction of a model by their sheer volume (Ananny & Crawford, 2018). Indeed, psychological work on causal explanation has shown that recipients prefer explanations involving a limited but relevant set of causes (Miller, 2019).

Usefulness

In addition to these two critiques of post-hoc explainability approaches, Rudin also questions their usefulness. In particular, she argues that post-hoc explainability techniques that produce heatmaps (e.g., Grad-CAM) are not very useful because they do not allow for good assessments of predictions. She illustrates this with an example: the most probable class (e.g., wolf) is often hardly distinguishable from that of a totally unrelated, less probable class (e.g., flute). While this may preclude the suitability of such explainability approaches for some applications, these heatmaps can at least suffice to detect if the model uses completely irrelevant features for the classification (see Lapuschkin et al., 2019 for an example). Furthermore, saliency maps are just one of a multitude of post-hoc explainability approaches that may be useful for different purposes. In addition to feature visualisation, we also mentioned TCAV above. TCAV can be used to query the influence of user-defined concepts (e.g., striped, female) on a classification, allowing for the detection of arbitrary biases in DNNs. Further, counterfactual explanations can be used to find out how to achieve a favourable decision from a black-box model (Karimi et al., 2021, 2022).

In summary, we think that there are post-hoc approaches that can be usefully employed to explain the output of black-box models even in high-stakes situations.

4.3.2 The limits of iterative feature engineering

One of Rudin's main claims is that, in most cases, there are ante-hoc explainable models that can achieve performance equivalent to black-box models. She suggests that through an iterative process of knowledge discovery and feature engineering, one can almost always arrive at such models. However, we believe that two obstacles stand in the way of such a process: hidden patterns and, relatedly, unengineerable features.

Hidden patterns

In a short but suggestive passage, Rudin proposes that one of the reasons for the widespread acceptance of the PET is a belief that opaque models exploit so-called “hidden patterns” in the data to achieve good performance (Rudin, 2019, p. 210). While Rudin does not explicitly define the term, we interpret it to refer to complex, non-linear, and/or distributed combinations of input features that capture predictively relevant

information in the input data. However, Rudin expresses scepticism about the notion that black-box models, such as DNNs, can uncover hidden patterns that cannot also be used by simpler models. She thinks reliance on these opaque models can be avoided given sufficient researcher ability in crafting the right kind of model. Rudin's discussion of hidden patterns is closely related to her other worries about black-box models. In particular, she argues that the assumption of hidden patterns is usually misleading and can give rise to false beliefs about the superior capabilities of opaque models. She believes that, often, instead of uncovering hidden patterns that reflect the genuine structure of the data domain in question, these models exploit spurious correlations in the training data, which may not necessarily reflect causality or any meaningful patterns in the real world.

Rudin raises valid concerns regarding the susceptibility of complex ML models to overfitting the training data, including noise and irrelevant features, resulting in suboptimal generalisation and unreliable predictions on new data (see Belkin et al., 2019; Yang et al., 2020 for discussions). However, recent theoretical work suggests that neural networks often do not overfit, even when their parameters outnumber the training samples (Zhang et al., 2021). More fundamentally, we think it is highly likely that hidden patterns do exist in some domains. Indeed, paying close attention to how and why Deep Learning (DL) became the dominant technique in ML gives us evidence for this.

In the early stages of ML, limited computational power and small datasets restricted the complexity of ML models it was possible to train (in terms of, e.g., the number of trainable parameters) (Goodfellow et al., 2016). Thus, most solutions relied on either manually crafted features or the application of general-purpose algorithms, such as (K)PCA, LDA, Fourier and Wavelet Transformation, SOM, KL Transform, Histograms, Clustering, and Embeddings, to automatically derive simple features (Jain et al., 2000; Webb, 2003). However, in highly complex domains, such as image recognition and natural language processing, both of these approaches failed to achieve adequate performance (LeCun et al., 2015). By efficiently leveraging computational resources to autonomously discover their own feature spaces during training, DL techniques have drastically reduced the need for human intervention in the feature engineering process (Goodfellow et al., 2016; LeCun et al., 2015; Schmidhuber, 2015). Given enough computational resources, the

data-driven approaches of highly parameterised ML models significantly outperform human engineering in various ML tasks (Chowdhery et al., 2023; Sutton, 2019). For one prominent example, consider the recent success of large language models. Increasing the parameter count of these models led to them learning the structure of natural language well enough to produce meaningful and coherent text (Brown et al., 2020; Chowdhery et al., 2023). We suggest that the most plausible explanation of this phenomenon is that increasing the scale of the models enabled them to learn subtle, complex dependencies that are difficult to articulate simply but are necessary for good performance on the difficult task of next-word prediction (Kaplan et al., 2020).

Feature spaces found by DL satisfy our criteria for hidden patterns: they are complex, non-linear, and use distributed representations. Further, they capture information that is likely required to solve the tasks at hand, as demonstrated by their performance advantage over (ante-hoc) explainable models. According to Rudin, even if there were hidden patterns, ante-hoc explainable models could also exploit them to achieve approximately equivalent performance. However, given these historical developments, this seems unlikely. The absence of satisfying results despite decades of research suggests simpler ML approaches may be incapable of learning the hidden patterns required for good performance in certain domains.

In sum, we think Rudin is right to consider that the question of when complex, opaque models are required is closely related to the question of which domains have “hidden patterns”. Indeed, we agree with Rudin's view that the assumption of hidden patterns and the mindless deployment of complex models can be harmful. However, we think her hope that there may be no domains containing hidden patterns which are too complex to be learned by simple models is unlikely to be true.

Unengineerable features

As outlined above, the transition from handcrafted to learned feature spaces has enabled improved performance on more complex and challenging problems (see also LeCun et al., 2015). The primary advantage of DL lies in its ability to perform well on tasks where it is difficult to encode human intuition using traditional algorithms or features. Consequently, DL has gained significant recognition in fields such as computer vision (e.g., AlexNet, see Krizhevsky et al., 2012; ResNet, see He et al., 2016), natural language

processing (e.g., Transformers, see Vaswani et al., 2017), and strategic game playing (e.g., AlphaGo, see Silver et al., 2017; AlphaStar, see Vinyals et al., 2019), where traditional algorithms tend to underperform.

Interestingly, DL outperforms handcrafted feature spaces even in domains where we already have a deep intuitive understanding of the problem (e.g., natural language processing). Many of the aforementioned tasks are so essential for humankind that almost every human has the innate ability to learn to perform reasonably well given some initial training. Nevertheless, encoding human knowledge and intuition into computer-readable formats, which usually requires a rigorous mathematical or logical representation, has proven difficult. This may be because features have to be built for detailed, low-level inputs like pixels in computer vision or tokens in natural language processing. Unlike such inputs, human reasoning happens at a higher-level (Marcus, 2001). Finally, even if some individual features could be extracted through a process of iterative feature engineering, the sheer quantity of such features may be too high to efficiently encode all of them by hand. This leads to a critical limitation in knowledge-discovery based approaches: even though we know what might constitute good features for a given ML task, we are sometimes unable to express them to computer systems.

4.3.3 The Rashomon set argument

The Rashomon set argument (Rudin, 2019; Semenova et al., 2022) proposes that simple (and thus explainable) models exist within the set of reasonably performant models (the Rashomon set) for almost every domain. The argument suggests that the reason simple models often exist is that the Rashomon set is large, containing many performant models. If there are many such models, then we should expect at least one of them to be ante-hoc explainable. By studying the Rashomon ratio, which measures the volume of performant models relative to the hypothesis space, one can estimate if a simple model is likely to exist before explicitly searching for it (Semenova et al., 2022). If many real-world datasets have large Rashomon sets, it implies that explainable models can, in principle, be used for important decisions without sacrificing performance. In essence, the Rashomon set argument provides a framework for understanding when and why performant and ante-hoc explainable models are likely to be found (Semenova et al., 2022).

The Rashomon set argument is certainly intriguing. However, despite being presented as speaking for the existence of ante-hoc explainable models across domains, the argument is weaker than it sounds. We give three reasons why.

Plausibility

The conclusion that Rudin draws depends upon two critical assumptions: i) that Rashomon sets will typically be large, and ii) that large Rashomon sets will contain simple models. However, neither of these assumptions are justified with rigorous argument. Accordingly, one can also reverse the claims: it may well be that the Rashomon set is not large, and, irrespective of this, it may be that the Rashomon set does not contain simple models. Although her analysis is interesting, Rudin has given us no independent reason to believe that explainable models will be contained in the Rashomon set in every domain.

Scope

We also note that Rudin conceives ante-hoc explainable models to refer to certain classes of models (e.g., decision trees). These models are deemed explainable because their features are usually hand-crafted and have a straightforward semantic interpretation. However, although these models may typically be explainable, that does not mean that they will remain explainable if they are allowed to grow arbitrarily large and represent numerous fine-grained distinctions (Lipton, 2018).

Time

Finally, even if the Rashomon set argument were valid, it would still not mean that finding ante-hoc explainable models would be easy. On the contrary, an existence proof does not mean that one will be able to find the model efficiently or even at all. While Rudin states that one could enumerate all performant models and thus eventually get to the explainable one, even finding optimal decision trees is an NP hard problem (Murthy, 1998), making it not guaranteed to get to a solution in a remotely reasonable time.

4.4 Re-Imagining the Performance-Explainability Trade-Off

Overall, though we find the thrust of Rudin's criticisms of the PET insightful, some of her claims are inadequately supported, unlikely to apply across domains, or make assumptions biased towards the conclusion she favours. In light of this, we sketch a more nuanced view that incorporates the valuable aspects of Rudin's criticisms but also accounts for its weaknesses. We start with expanding the PET.

4.4.1 From PET to PET+: Incorporating further factors

The discussion of the Rashomon set argument revealed that, even if it were true in theory, one would be required to expend a great deal of time enumerating and evaluating models in order to obtain an explainable and performant one. This is not the only point in Rudin's argument where development resources come into play. Her proposal of an iterative feature engineering process also demands a significant time investment as well as researcher expertise and supporting resources (e.g., tools, money, etc.).

However, when Rudin considers the use of black-box models explained with post-hoc techniques, she does not account for variation in this additional dimension. In particular, she does not consider how well black-box models could be explained given development resources comparable to those required for iterative feature engineering (e.g., by systematically reverse engineering complex models with model-specific explainability techniques). To compare black-box methods and iterative feature engineering fairly, however, we need to consider what these two approaches can achieve given the same amount of resources. We think this is a crucial point and suggest that the PET needs to be refined to accommodate the availability of resources.

Thus, we suggest that the availability of developmental resources, especially time, constitutes an extra dimension for the PET. Making the simplifying assumption that expertise and financial budget are held fixed, we attain a three-factor trade-off of time, performance, and explainability: PET+.

FIGURE 11: AN IDEALISED DEPICTION OF PET+

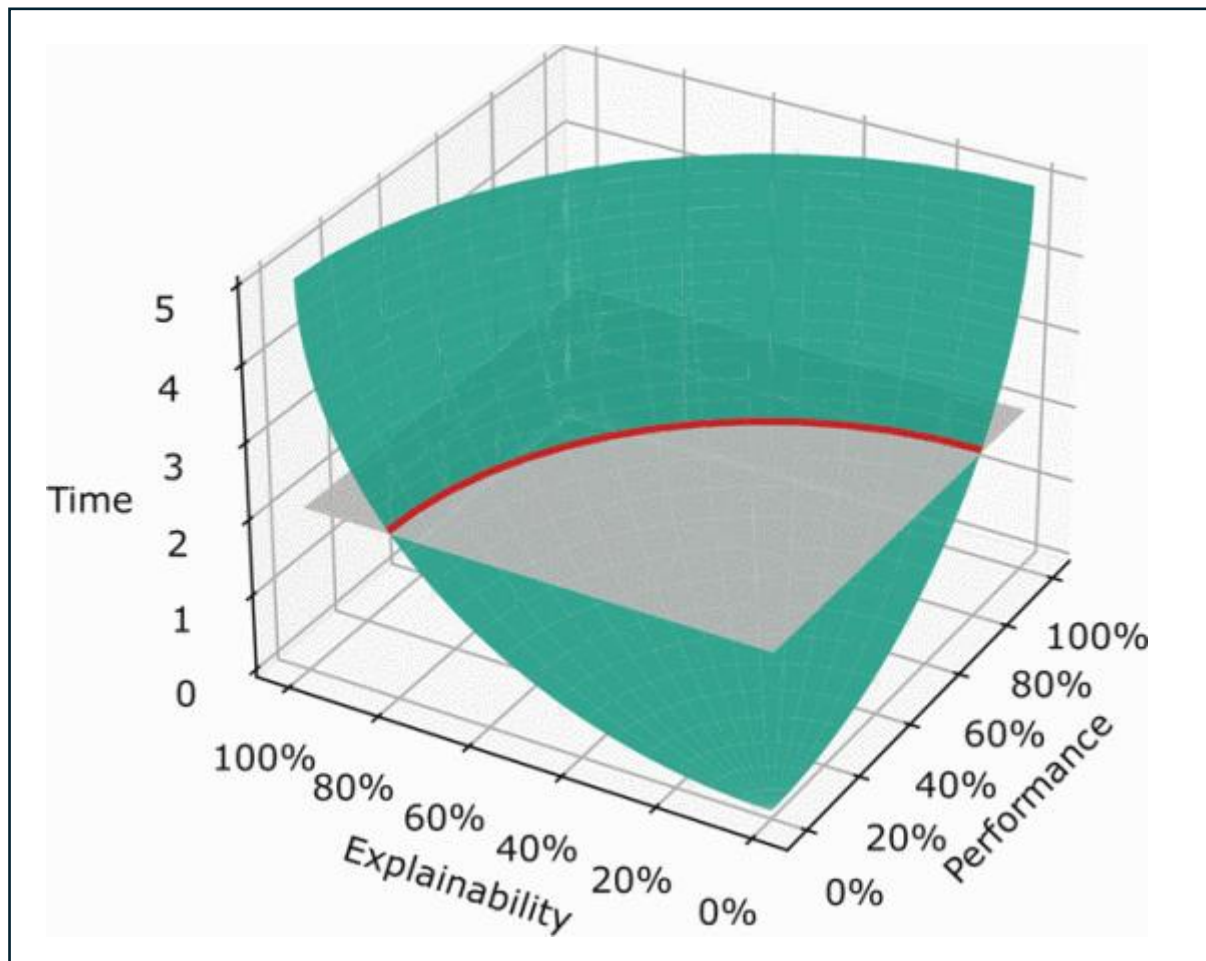


Figure 11 shows an idealised depiction of PET+ as a trade-off with the three dimensions *explainability*, *performance*, and *time*. With more time available, better solutions can be found. For a given temporal budget (e.g., grey plane), however, a choice has to be made between the solutions of the Pareto set (red curve).

(PET+) *For some real-world tasks, when developing a model to perform (part of) the task, increasing the performance of the model necessarily decreases the explainability of the model (and vice versa) given the same temporal resources. By increasing the temporal budget, we may achieve a Pareto improvement (Ehrgott, 2005) on our previous best model (see Fig. 11).*

We suggest that, in many cases, developing explainable models that are also highly performant requires extensive development resources like time, suitably experienced researchers, and special tools. In general, we agree with Rudin's claim that high-stakes domains justify the deployment of such resources. As such, we think PET+ is a more

appropriate characterisation of the issue that (requirements) engineers are facing than PET. However, we think this holds true for *both* Rudin's preferred method of designing ante-hoc explainable models and approaches which rely on training black-box models and then explaining them afterwards. Thus, it is crucial to apply this consideration equally to *all* possible approaches to engineering ML models. Only by doing this can the strengths and weaknesses of different approaches be accurately compared.

Finally, we note that the three-factor trade-off captured by PET+ still does not describe the full complexity of model development. Due to Rudin's focus on high-stake situations, she discourages the use of explanations lacking fidelity and appeals for the use of explainable models partially for *moral* reasons. We agree with Rudin that moral considerations should come into play when deciding which values to prioritise in model-building. Indeed, we think that moral risks should be considered alongside other risks that developers take on such as financial risks and risks to one's reputation.

FIGURE 12: FACTORS INFLUENCING PET+

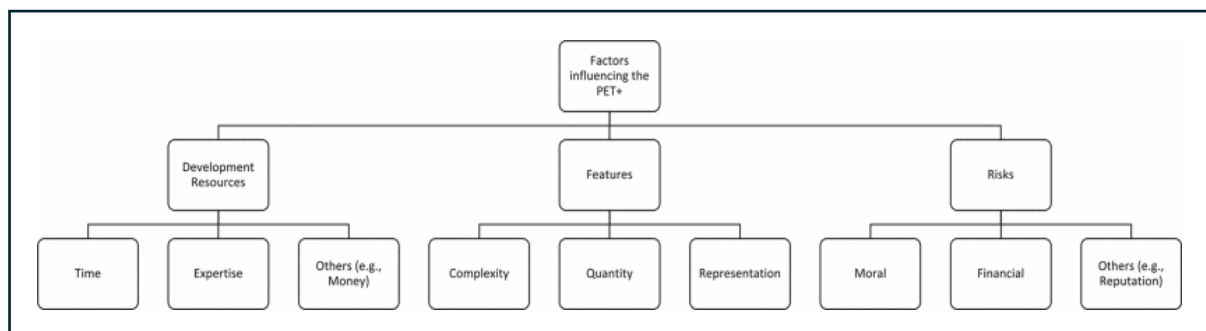


Figure 12 depicts our classification of the various factors that influence the PET+. The levels of performance and explainability it is possible to achieve are constrained by the development resources (e.g., time, expertise) available and the properties of the features (i.e., their complexity, quantity, and representation) that exist in the particular domain. Risks (e.g., moral, financial) affect the relative importance of performance and explainability in any given real-world scenario.

4.4.2 PET+ as a multi-objective optimisation problem

Overall, the performance-explainability trade-off is influenced by many factors (e.g., development resources, domain details, and risks; see Fig. 12). Further, which outcomes are considered most desirable in each real-world scenario depends critically on the relative importance we assign to performance and explainability. To decide how to

proceed in light of this, it will be beneficial to have a framework that incorporates all of the relevant factors. To this end, we suggest casting PET+ as a multi-objective optimisation problem as follows.

Consider a specific decision (e.g., making medical diagnoses) which, at $t=0$, is made by a system (which may be a human) that produces some cost at every time-step. The cost is inversely proportional to a linear combination of performance and explainability (i.e., the better the performance and the greater the explainability of the model, the lower the cost). With this setup in mind, researchers aim to reduce the cost by evaluating and selecting a sequence of actions over time (i.e., developing, refining, explaining, and ultimately, deploying new models). To choose which actions to take, they need to estimate the value (i.e., reduced cost relative to some baseline) of different sequences of actions based on their expertise and domain knowledge. This value will be proportional to the improvements to performance and explainability they can produce, as well as how quickly they can produce those improvements. The researchers' ultimate goal is to minimise the *level of regret*, defined as the difference between the accumulated cost-over-time and the cost that would be incurred by the best sequence of actions.

Each of the factors we have discussed is captured by this formulation of the problem. First, consider *development resources*. The specific *expertise* researchers bring to a problem constrains what level of performance and explainability they might achieve. In our framework, researcher skills can be thought of as restrictions on the set of actions available. Similarly, researcher knowledge (e.g., about the nature of the relationship between symptoms and various medical conditions) constrains how accurately the value of different sequences of actions can be calculated. At the same time, temporal constraints limit the length of action sequences that are possible, and, thereby, the level of performance and explainability it is possible to achieve.

Next, consider what kind of *features* are present in the domain. If the features which must be learned to achieve good performance are complex, numerous, and best represented in a distributed way (i.e., there are *hidden patterns*), it may be that constructing black-box models and then explaining them post-hoc is the best approach. Where they are absent, an iterative feature engineering approach may be preferable. In our framework, these

feature-based considerations constrain the true value of different actions (i.e., how much they actually improve performance and explainability).

Finally, *risk* enters the framework in the relative weighting of performance and explainability. Depending on which values are deemed to be most important in the real-world scenario in question, the weights will be set differently, thus altering how the cost is calculated. For example, explainability may be particularly important in a judicial setting (Eliot, 2021), while performance might be prioritised in a medical scenario (London, 2019).

Practically, we think that model development, in light of the PET+, can be usefully informed by asking the following heuristic questions: What expertise and tools are available? What is the temporal and financial budget? What properties does the data domain in question have? How important are performance and explainability relative to one another? Given a set of answers to these questions, one should consider which actions to take to minimise expected regret.

4.5 Conclusion

Our reflections on the performance-explainability trade-off are intended to serve as a catalyst for further discussion and exploration within the RE community. We believe that by painting a more nuanced picture of the trade-off, captured by PET+, we can inspire researchers, practitioners, and stakeholders to collaborate in shaping the future of RE for AI systems.

4.6 References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems*, 31. https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7e7fea049b8737-Abstract.html
- Amorim, J. P., Abreu, P. H., Santos, J., Cortes, M., & Vila, V. (2023). Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, 60(2), 103225. <https://doi.org/10.1016/j.ipm.2022.103225>

- Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, 7, e479. <https://doi.org/10.7717/peerj-cs.479>
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, 35(1), 12. <https://doi.org/10.1007/s13347-022-00510-w>
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849–15854. <https://doi.org/10.1073/pnas.1903070116>
- Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It’s Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–266. <https://doi.org/10.1145/3531146.3533090>
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. <https://doi.org/10.1109/tpami.2013.50>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>
- Brunotte, W., Chazette, L., Klös, V., & Speith, T. (2022). Quo Vadis, Explainability? – A Research Roadmap for Explainability Engineering. In V. Gervasi & A. Vogelsang (Eds.), *Requirements Engineering: Foundation for Software Quality* (Vol. 13216, pp. 26–32). Springer International Publishing. https://doi.org/10.1007/978-3-030-98464-9_3
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. <https://doi.org/10.48550/arXiv.2303.12712>
- Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, 5(3), 10.23915/distill.00024. <https://doi.org/10.23915/distill.00024>

- Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. *2021 IEEE 29th International Requirements Engineering Conference (RE)*, 197–208. <https://doi.org/10.1109/RE51729.2021.00025>
- Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: Challenges and recommendations. *Requirements Engineering*, 25(4), 493–514. <https://doi.org/10.1007/s00766-020-00333-1>
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240), 1–113.
- Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32. <https://proceedings.neurips.cc/paper/2019/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html>
- Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. <https://doi.org/10.23919/MIPRO.2018.8400040>
- Ehrgott, M. (2005). *Multicriteria Optimization*. Springer Science & Business Media.
- Eliot, D. L. B. (2021). *The Need For Explainable AI (XAI) Is Especially Crucial In The Law* (SSRN Scholarly Paper 3975778). Social Science Research Network. <https://doi.org/10.2139/ssrn.3975778>
- Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1915–1929. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.2012.231>
- Freitas, A. A. (2014). Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1), 1–10. <https://doi.org/10.1145/2594473.2594475>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Gosiewska, A., Kozak, A., & Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, 150, 113556. <https://doi.org/10.1016/j.dss.2021.113556>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 93:1-93:42. <https://doi.org/10.1145/3236009>
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- Ha, S., & Jeong, H. (2021). Unraveling hidden interactions in complex systems with deep learning. *Scientific Reports*, 11(1), Article 1. <https://doi.org/10.1038/s41598-021-91878-w>

- He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. 770–778.
https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
- Herzog, C. (2022). On the Ethical and Epistemological Utility of Explicable AI in Medicine. *Philosophy & Technology*, 35(2), 50. <https://doi.org/10.1007/s13347-022-00546-y>
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6), 82–97. *IEEE Signal Processing Magazine*.
<https://doi.org/10.1109/MSP.2012.2205597>
- Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1), 141–154.
<https://doi.org/10.1016/j.dss.2010.12.003>
- Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4–37. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
<https://doi.org/10.1109/34.824819>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361). arXiv. <https://doi.org/10.48550/arXiv.2001.08361>
- Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.*, 55(5), 95:1-95:29. <https://doi.org/10.1145/3527848>
- Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic Recourse: From Counterfactual Explanations to Interventions. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362.
<https://doi.org/10.1145/3442188.3445899>
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 2668–2677. <https://proceedings.mlr.press/v80/kim18d.html>
- Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368.
<https://doi.org/10.1109/RE.2019.00046>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25.
https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), Article 1. <https://doi.org/10.1038/s41467-019-08987-4>
- Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1), 98–113. *IEEE Transactions on Neural Networks*. <https://doi.org/10.1109/72.554195>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. <https://doi.org/10.1038/nature14539>
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. *Proceedings of the IEEE*. <https://doi.org/10.1109/5.726791>
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Liu, H., & Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Springer US. <https://doi.org/10.1007/978-1-4615-5689-3>
- London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1), 15–21. <https://doi.org/10.1002/hast.973>
- Maimon, O., & Rokach, L. (Eds.). (2005). *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag. <https://doi.org/10.1007/b107408>
- Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023). Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy. *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 337–342. <https://doi.org/10.1109/REW57809.2023.00063>
- Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. <https://doi.org/10.1145/3287560.3287574>
- Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2(4), 345–389. <https://doi.org/10.1023/A:1009744630224>

- Nguyen, A., Yosinski, J., & Clune, J. (2019). Understanding Neural Networks via Feature Visualization: A Survey. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 55–76). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_4
- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, 2(11), e7. <https://doi.org/10.23915/distill.00007>
- Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010. <https://doi.org/10.23915/distill.00010>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer International Publishing. <https://doi.org/10.1007/978-3-030-28954-6>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2), 336–359. <https://doi.org/10.1007/s11263-019-01228-7>
- Semenova, L., Rudin, C., & Parr, R. (2022). On the Existence of Simpler Machine Learning Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1827–1858. <https://doi.org/10.1145/3531146.3533232>
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), Article 7587. <https://doi.org/10.1038/nature16961>
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359. <https://doi.org/10.1038/nature24270>
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps* (arXiv:1312.6034). arXiv. <https://doi.org/10.48550/arXiv.1312.6034>
- Speith, T. (2022a). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. <https://doi.org/10.1145/3531146.3534639>
- Speith, T. (2022b). How to Evaluate Explainability? - A Case for Three Criteria. *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, 92–97. <https://doi.org/10.1109/REW56159.2022.00024>

- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). *Striving for Simplicity: The All Convolutional Net* (arXiv:1412.6806). arXiv. <https://doi.org/10.48550/arXiv.1412.6806>
- Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1), 109–133. <https://doi.org/10.1093/bjps/axz035>
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3104–3112.
- Sutton, R. (2019). *The bitter lesson*. <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>.
- Tschantz, M. C. (2022). What is Proxy Discrimination? *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1993–2003. <https://doi.org/10.1145/3531146.3533242>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. <https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., ... Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350–354. <https://doi.org/10.1038/s41586-019-1724-z>
- Webb, A. R. (2003). *Statistical Pattern Recognition*. John Wiley & Sons.
- Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking Bias-Variance Trade-off for Generalization of Neural Networks. *Proceedings of the 37th International Conference on Machine Learning*, 10767–10777. <https://proceedings.mlr.press/v119/yang20j.html>
- Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288. <https://doi.org/10.1007/s13347-019-00382-7>
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 107–115. <https://doi.org/10.1145/3446776>
- Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5), Article 5. <https://doi.org/10.3390/electronics10050593>

© [2023] IEEE. Reprinted, with permission, from: Barnaby Crook, Maximilian Schlüter, & Timo Speith. Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI). 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW). September 2023.

Chapter 5:

Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience

This chapter is forthcoming as “Crook, B., & Kästner, L. (Forthcoming). Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Springer.” Slight changes have been made to fit the dissertation format.

5.0 Abstract

Machine learning models, particularly deep artificial neural networks (ANNs), are becoming increasingly influential in modern neuroscience, often being employed as models of cognitive and perceptual systems. However, these models are often complex and opaque, leading some to worry that, by utilising ANNs, neuroscientists are trading one black box for another. On this view, despite increased predictive power, ANNs effectively hinder our scientific understanding of the brain. We think these worries are unfounded. While ANNs are difficult to understand, there is no fundamental trade-off between the predictive success of a model and how much understanding it can confer. Thus, utilising complex computational models in neuroscience will not generally inhibit our ability to understand the (human) brain. Rather, we believe, deep learning is best conceived as offering a novel and unique epistemic perspective for neuroscience. As such, it affords insights into the operation of complex systems that are otherwise unavailable. Integrating these insights with those generated by traditional neuroscience methodologies bears the potential to propel the field forward.

5.1 Introduction

Machine learning in general, and deep artificial neural networks (ANNs) in particular, have become an increasingly influential methodological tool in the arsenal of modern computational neuroscience. Because of their flexibility, ANNs are used in multiple distinct ways, from the identification of predictive variables to serving as models of cognitive and perceptual faculties (Glaser et al., 2019). As a result of their expressive power, ANNs have achieved unmatched success when it comes to task performance, prediction of neural activity, and replication of human-like behavioural patterns (Cao & Yamins, 2021; Kanwisher et al., 2023; Schrimpf et al., 2020). However, despite their undeniable predictive power, deep learning-based approaches to investigating the brain are not without limitations. One such limitation is that systems designed through deep learning are often highly complex and opaque to researchers. Because their functional structure is generated through an automated learning procedure, precisely *how* deep learning-based systems solve a given set of computational problems often remains unclear, even to their designers (Creel, 2020; Zednik, 2021; Samek & Müller, 2019). This applies not only to the use of deep learning in neuroscience but to deep learning systems at large. Many AI researchers worry that there is a “tension between machine learning performance (predictive accuracy) and explainability” (Gunning & Aha, 2019, p. 45; see Crook et al., 2023 for discussion). This general worry is also echoed with respect to neuroscience research specifically. For instance, Lindsay argues that “to have a model complex enough to perform real-world tasks, we must sacrifice the desire to make simple statements about how each stage of it works” (2021, p. 2024). Building on such worries, Chirimuuta portrays the situation as one where computational neuroscience faces “a trade-off between predictive accuracy and the ability of [its] models to confer understanding” (2021, p. 787).

While we agree with Lindsay's statement, we believe that the fear expressed in Chirimuuta's is misguided. There might indeed be a trade-off between the predictive success of individual models and the ease of understanding them; but this does not imply that utilising complex computational models in neuroscience will generally inhibit our ability to understand the (human) brain. On the contrary: We firmly believe that rather than worrying about sacrificing understanding by utilising deep learning in neuroscience,

we should welcome ANNs into the neuroscientists' toolkit.⁸⁸ They bear the potential to generate new insights about the brain from a yet unavailable *epistemic perspective* (see Kästner, 2018) that is complementary to the perspectives provided by established neuroscientific methodologies. That is, deep learning contributes to, and does not detract from, understanding neuroscientific phenomena.

We shall begin our exposition by diagnosing the source of trade-off worries in a misconception of scientific understanding (Chapter 5.2). The reasoning behind the notion that researchers are paying a high price for predictive success crucially depends on weighing models' predictive power against their intelligibility. While it is widely agreed that the aim of science is to understand *phenomena*,⁸⁹ the argument for a trade-off depends on equating this understanding with the intelligibility of particular models of those phenomena (see, e.g., Bokulich, 2017; Elgin, 2017).⁹⁰ However, we think this is mistaken in two ways. First, while we agree that models play a key role in the production of scientific understanding, we think such understanding depends upon integration of numerous distributed sources of explanatory information. Second, the intelligibility of a model is not a reliable guide for how much understanding it can ultimately confer. While interpreting complex models may be challenging, so long as it is feasible, even relatively unintelligible models can induce scientific understanding of phenomena.

Once we recognise this distinction between model understanding and *model-induced understanding of phenomena*, it becomes evident that the prediction-understanding trade-off is a bogeyman we need not fear: Deep learning can contribute to a progressive computational neuroscience that yields both improved predictive performance *and* deeper scientific understanding (Chapter 5.3). This can be achieved by integrating insights from traditional neuroscientific research with those delivered by ANNs and systematic post-hoc interpretability strategies. Insights from these different epistemic

⁸⁸ For ease of reading, we shall use the terms “deep learning models” and “ANNs” synonymously.

⁸⁹ There is a rich debate in contemporary philosophy of science about what precisely phenomena are and how to best characterise them (e.g., Colaco, 2019; Craver & Darden, 2013; Feest, 2017; Shagrir & Bechtel, 2017). Here we simply use the term “phenomenon” to refer to any explanandum investigated by neuroscientists. Prototypical examples include such things as color vision, face recognition, or cortical language processing.

⁹⁰ We use the term *intelligible* to capture how easy it is to understand a given model.

perspectives will contribute to a store of distributed knowledge that neuroscientists can draw on to explain and understand phenomena.

Because debates about what it takes to explain and understand complex systems have a particularly rich history in neuroscience, we focus our discussion on this field. Still, we think our insights will apply not only to neuroscience; but are likely to generalise to relevantly similar domains where deep learning is being utilised to study complex systems.

5.2 Trade-Off Worries

In this section, we discuss why trade-off worries come about. We shall first review arguments to the effect that a trade-off exists between the predictive success of a model and how much understanding it can confer (Chapter 5.2.1). Against this backdrop, we shall explain why we think these arguments are flawed: because they mistake model understanding for model-induced understanding of phenomena (Chapter 5.2.2).

5.2.1 Why worry?

A common conception is that models inducing understanding should be simple and intelligible, abstracting away from or usefully idealising the complexities of the real world (cf., Bokulich, 2017; Elgin, 2017; Potochnik, 2016). Against this backdrop, it seems natural to expect that scientific understanding is best conferred by simple transparent models that domain experts understand perfectly. In traditional scientific modelling, scientists' domain knowledge is encoded by the construction of a model of a target phenomenon (Craver, 2001; Frigg & Hartmann, 2020). These models can frequently be represented by intuitive box-and-arrow diagrams or simple equations experts can grasp just by eyeballing them. The components of these models are often taken to represent relevant features of target phenomena, though these may be individuated functionally rather than structurally (Bennett et al., 2019; Craver & Kaplan, 2020). The construction, evaluation, and manipulation of models that represent phenomena is often taken to be central to the production of scientific understanding (de Regt, 2009; Elgin, 2007; Morgan & Morrison, 1999).

Deep learning models are different on various scores.⁹¹ First, because they are trained with large datasets and consist of many parameters, they can be exceedingly complex. As such, they break with the expectation of being simple. Second, and relatedly, unlike with traditional modelling, domain knowledge typically only constrains choices about the initial *architecture* and training data of deep learning-based models (Goodfellow et al., 2016). The architecture consists of the number of layers, the number and types of neurons in each layer, connectivity profiles of each neuron, and learning rules of the network (Chollet, 2021; Richards et al., 2019). While these features are specified by researchers, the rest of what a trained ANN model encodes (e.g., what representations the system learns and how its functional units interact to solve the computational problem) is learned through an automated training procedure, typically involving backpropagation and some form of gradient descent (LeCun et al., 2015). On the one hand, this autonomy from human design is the central strength of deep learning. It is what gives ANNs the flexibility to learn subtle, task-relevant patterns that elude traditional modelling and thus to expand the range of domains in which models can achieve predictive success (Boge, 2022; Jumper et al., 2021; Zhuang et al., 2021). On the other hand, especially when ANNs have many parameters, it is often unclear both i) how the internal structure of a trained model implements the mapping from input to output (see Chirimuuta, 2021; Creel, 2020), and ii) what relation this internal structure bears to the target phenomenon (this is what Sullivan, 2022 calls 'link uncertainty').⁹² As such, deep learning models can be predictively successful without necessarily illuminating how exactly the target phenomenon is produced; that is, without conferring understanding (See Chapters 8 and 9 in Durán & Pozzi (Eds.), forthcoming).

As a result of these unique properties, several authors writing about the increasing prevalence of deep learning in science have worried that the advantages we accrue in

⁹¹ The word 'model' has multiple meanings. In statistics, any deep learning system is called a model, regardless of the purpose for which it is constructed. However, in the philosophy of science, a model is any structure, mathematical, physical, or graphical, that is constructed to represent (some aspect of) a target system. In neuroscience, statistical models based on deep learning are used both as tools (e.g., for data pre-processing; cf. Glaser et al., 2019) *and* as models that represent target systems (Cichy & Kaiser, 2019). In what follows, we focus on the use of ANNs as *scientific* models (i.e., models constructed to represent).

⁹² There is a vast body of literature on opacity. For a classic see Burrell (2016), for a recent comprehensive taxonomy see Mann et al. (2023), for further discussion see Chapters 1 and 2 in Durán & Pozzi (Eds.), forthcoming.

terms of predictive accuracy are offset by a decrease in explanatory power and understanding (Boon, 2020; Chirimuuta, 2021; Lindsay, 2021; López-Rubio & Ratti, 2021; Srećković et al., 2022). For example, Srećković and colleagues suggest that increased use of deep learning models may “lead us away from the explanatory aspect of science” (2022, p. 172). Along similar lines, Boon imagines a future in which “scientific researchers and scientific knowledge become superfluous as learning from large data sets, algorithms and data-models will be developed at a degree of complexity and adequacy far beyond the capacity of the human intellect” (2020, pp. 49–50). Finally, López-Rubio and Ratti state that, “when predictive performances increase, the possibility of elaborating a (mechanistic) explanation necessarily decreases” (2021, p. 1332). While the exact claims and arguments of these authors differ slightly, the overall perspective they present is clear: we must choose between increasing the predictive adequacy of our models and gaining understanding from them; we cannot have both.

Worries about this hypothesised trade-off between predictive success and understanding have touched many areas of science, from protein-folding (Bouatta et al., 2021), to physics (Boge, 2022; see also Chapter 19 in Durán & Pozzi (Eds.), forthcoming), to molecular biology (López-Rubio & Ratti, 2021). In what follows we concentrate on neuroscience, the discipline with which we are most familiar. Though we spell out our response with this specific focus, we suggest that the broad shape of our argument is likely to carry over to relevantly similar scientific domains, even if the details may differ.

Perhaps due to the intertwined history of neuroscience and AI research (Hassabis et al., 2017; McCulloch & Pitts, 1943), deep learning has had a particularly large and rapid influence on the study of the brain and nervous system (Cichy & Kaiser, 2019; Marblestone et al., 2016; Richards et al., 2019). As well as being used to identify predictive features and solve engineering problems (Glaser et al., 2019), deep learning has been employed in computational neuroscience to produce ANN models of cognitive and perceptual systems (Doerig et al., 2023; Kell et al., 2018; Schrimpf et al., 2020; Yamins & DiCarlo, 2016). It is in this capacity that deep learning's contributions to neuroscience have prompted worries of a trade-off. While ANNs have proven successful at implementing solutions to ethologically relevant tasks, predicting neural activity, and recapitulating human-like error profiles, critics have raised concerns that we are simply

replacing one black box with another (Charles Leek et al., 2022; Chirimuuta, 2021; Thompson, 2021). This argument has been made most explicitly by Mazviita Chirimuuta. We shall briefly reiterate her argument before explaining why we do not agree with its conclusion.

Chirimuuta directs her argument at models “intended to represent the functions computed by neural systems” (2021, p. 772).⁹³ As she details, traditional models, such as Georgopoulos and colleagues’ (1986) *Population Vector Algorithm* for estimating arm movements from primary motor cortex activity, are fairly simple. Although such models knowingly encode false assumptions, in Chirimuuta’s estimation they end up being “highly intelligible, representationally inaccurate but surprisingly useful” (2021, p. 774). At the other end of the spectrum, models based on deep learning, such as Sussillo and colleagues’ (2012) recurrent neural network for decoding arm kinematics from neural activity, are complicated beasts. As Chirimuuta puts it, such models are “non-linear, opaque but predictively accurate” (2021, p. 776). This contrast recapitulates the distinction between traditional modelling strategies and deep learning that we have already discussed; and it leads directly to the suggestion that “there is a trade-off between a model’s predictive power and its ability to increase the scientist’s understanding of a neural response” (Chirimuuta, 2021, p. 768). However, we think this is too quick. To reach this conclusion, Chirimuuta uses the notion of ‘model intelligibility’ to capture not just the idea that models are easy to understand but *also* their overall ability to confer understanding. It is the conflation of these two notions that we take issue with.⁹⁴

⁹³ Chirimuuta argues for a principled distinction between *efficient coding* models which aim to capture the encoding functions used by the brain and *mechanistic* models which (also) encode anatomical isomorphisms, intending her argument to apply only to the former. However, in our view, the properties of ANNs that Chirimuuta invokes to make her case creates a structurally analogous problem for both kinds of model. As such, we take our response to address a generalised version of Chirimuuta’s argument. This argument covers a broad class of ANN models in neuroscience, including those with mechanistic explanatory goals.

⁹⁴ We wish to stress that this is not a semantic dispute about how *intelligibility* should be defined, but a conceptual dispute about whether ‘ease of understanding’ and ‘ability to confer understanding’ ought to be equated with one another.

At this point, it is important to introduce a subtle but important distinction concerning the source of the unintelligibility or opacity of ANNs.⁹⁵ On the one hand, ANNs are trained via automated learning procedures. This is what makes it possible for researchers to develop models that solve certain tasks without understanding exactly how they do so. On the other hand, ANNs are typically very complex, consisting of many tunable parameters which interact in subtle ways to produce the behaviour of the whole system. While these two sources of opacity are intimately related in the methodology of deep learning, they are conceptually distinct. The first problem is that an explanation is not *given* (i.e., automatically conferred through construction of the model), the second is that an explanation is not *simple*. In principle, the first source of opacity might be overcome through systematic post-hoc analysis and interpretation of a trained ANN (as we will discuss in Chapter 5.3). Assuming we might reverse engineer the functional structure of ANNs, however, we are still left with the second source of opacity, viz. the structure we uncover might be quite complex. Chirimuuta states clearly that in her view this second source is sufficient to ensure the trade-off obtains:

“[...] even if we could write down by hand the equations embedded in the trained ANN's [...], those models would still be far less intelligible than their low-tech predecessors, because they would be nonlinear and contain very many more terms than the ones occurring in the traditional models.” (2021, p. 782)

So, even if an explanation was given, it would not necessarily alleviate opacity because it might lack the simplicity required for humans to understand what a system is doing. In other words, even sophisticated interpretability methods may be powerless to overcome the trade-off between the predictive success of deep learning models and their potential to confer understanding. With these arguments in view, let us turn to our own assessment of the trade-off story.

⁹⁵ We take opacity to mean “barrier to conferring understanding”. Notice that opacity is at least partially a subjective phenomenon. What is opaque to a layperson may not be opaque to an expert (Humphreys, 2009).

5.2.2 Why not to worry!

We grant that there might be a trade-off between predictive success and ease of understanding for individual models. However, we take issue with the notion that this implies neuroscientists necessarily sacrifice understanding by utilising ANNs in their research. This is because what they are after, we think, is not intuitive understanding of models of complex systems but understanding the *phenomena* that these systems exhibit with the help of models. Thus, there is no requirement for models to be simple. What matters is that they confer *understanding of the target phenomenon*, even if that understanding requires significant interpretative work. Indeed, as we will show below, there is nothing in principle stopping complex models from conferring greater understanding than simple ones.

Before we continue, a note on what we take (scientific) understanding to be is in order. Generally speaking, we are sympathetic to Elgin's suggestion that "understanding is a grasp of a comprehensive body of information that is grounded in fact, is duly responsive to evidence, and enables non-trivial inference, argument, and perhaps action regarding that subject the information pertains to" (2007, p. 39). Put simply, we take understanding to be having some kind of *qualitative grasp* of a target phenomenon (de Regt, 2009; Elgin, 2007; Strevens, 2013), some kind of *insight* into how a phenomenon of interest is exhibited by a system. It is a widely acknowledged feature of grasping that it is gradable, i.e., it comes in degrees (Elgin, 2007; Hills, 2016). With this in mind, we can say a subject S grasps a phenomenon P to the extent that they are able to (i) make coarse-grained predictions about P without performing exact calculations (cf., de Regt, 2009), (ii) give rich, detailed, and productive answers to a broad range of questions concerning P (cf., Chirimuuta, 2021), and (iii) intervene on and control P (or at least to describe such interventions, even if they would be impractical) (cf., de Regt, 2009). To serve these purposes, S might utilise some kind of (mental) representation of P where this representation might be considered an explanation for P or, alternatively, enable S to explain P.⁹⁶ Note that (i)-(iii) are also features commonly associated with good scientific explanations (see e.g., Craver, 2007; Pearl, 2009; Salmon, 1984; Woodward, 2003). This

⁹⁶ Given grasping is a gradable notion, (i)-(iii) can be performed with various degrees of correctness or success.

tight connection between understanding and explanation is no accident. If Alice has a better grasp of P than Bob (in the sense of (i)-(iii)), then she is likely to not only have a better understanding of P but to also be able to *explain* it better.

Admittedly, this characterisation of understanding in terms of grasping is fairly minimal. What we say about the relation between (i), (ii), and (iii), about the role of (mental) representations, and about the precise relation between understanding and explanation remains intentionally sketchy. But since the argument we make in what follows is compatible with a range of more articulated accounts of understanding (see de Regt, 2009; Elgin, 2007; Hills, 2016; Strevens, 2013 for various detailed views; see Grimm, 2021 for an overview), we think this is a merit rather than a shortcoming. What is important for the current discussion is merely that scientific understanding is conferred by the very kind of information we utilise in scientific explanations. And that the targets of understanding and explanation in science are *phenomena* (e.g., Bechtel & Richardson, 2010; Colaco, 2019; Craver, 2007; Feest, 2017; Glennan, 2017; Hochstein, 2016; Kästner, 2018).

Let us now turn to our assessment of the trade-off arguments presented above (Chapter 5.2.1). In those arguments, a trade-off between model complexity and intelligibility is taken to imply a trade-off between model complexity and the potential of a model to confer understanding of phenomena. But understanding models is not the same as understanding phenomena. And the ease of understanding a model is not the same as how much understanding that model can confer. Granted, all of the tell-tale signs of grasping just described can be straightforwardly applied to models and their behaviour (i.e., by casting them as target phenomena). However, while models play a crucial instrumental role in our understanding of phenomena, they are not typically the ultimate *targets* of inquiry in neuroscience or any other natural science; models are usually *mediators* (cf., Morgan & Morrison, 1999), not explananda. Thus, we must take care not to mistake model understanding for phenomenon understanding.⁹⁷

⁹⁷ This is not to say model understanding cannot contribute to understanding phenomena. Indeed, model understanding plays a crucial role. However, the relationship between understanding models and understanding phenomena is complex. We will elaborate on this in Chapter 5.3.2.

The issue we are getting at is nicely captured by Lawler and Sullivan's (2021) distinction between *model explanation* and *model-induced explanation*.⁹⁸ In model explanation, the content of the model simply is the explanation. In model-induced explanation by contrast, "the relevant explanatory information is independent of the model but only closely intertwined with the model due to the history of obtaining the information" (Lawler & Sullivan, 2021, p. 1069). Talking about understanding we can reframe the point thusly: Whenever our understanding of a phenomenon includes information extrinsic to any particular model, model-induced understanding of a phenomenon is not reducible to model understanding. We think this is actually very commonplace in scientific practice (cf., Hochstein, 2016; Kästner, 2018; Mitchell, 2002, 2019). Thus, reducing understanding of phenomena to what is encoded in a model is both descriptively and normatively inadequate. It is descriptively inadequate because real-world scientific understanding of complex phenomena frequently depends upon facts and details extrinsic to any individual model. It is normatively inadequate because focusing on understanding models – rather than model-induced understanding – artificially prevents us from satisfying scientific desiderata that might otherwise be satisfied (i.e., increasing both predictive accuracy *and* understanding of phenomena).

With these distinctions in mind let us return to Chirimuuta's argument. Recall her claim that "models of neural systems are either very intelligible, or predictively accurate, but not both" (2020, p. 781). In a certain sense, we agree with this claim. It is obviously *easier* to understand simple models than it is to understand complex models (perhaps tautologically so). However, that does not imply that we obtain a greater understanding of phenomena through constructing and using simple models than we do through constructing, using, and interpreting complex ones. In fact, we believe complex models will often confer higher degrees of understanding of phenomena in the sense described above: They enable us to (i) make better predictions, (ii) answer more questions about a phenomenon, and (iii) allow for more intervention and control.

⁹⁸ Note that Lawler and Sullivan are talking about explanation rather than understanding. As we stated above, we see an intimate link between the two. Depending on how exactly the relation between understanding and explanation is construed, the distinction might even collapse (e.g., when understanding is conceptualised as the possession of a mental representation of the right kind of explanation) (see Strevens, 2013 for an example of such a view).

But what about the opacity of ANNs – does it not interfere with (i)-(iii)? To address this question, recall our distinction between whether explanatory information encoded in an ANN is *given* and whether it is *simple* (Chapter 5.2.1). Suppose that the *givenness* issue is overcome and we obtain an equation describing the function the model has learned, e.g., through interpretability methods (we will address the feasibility of this in Chapter 5.3.3). Chirumuuta argues that, even so, “eyeballing an equation of such complexity would not give the neuroscientist the same qualitative sense of how adjustment of parameters or variables would make a difference to the behaviour of the system [as traditional, simpler models]” (2021, p. 782). Here, we think Chirumuuta conflates two things: the ease of coming to understand something is not the same as the degree of understanding that one might obtain from it. To make the point clear, recall Lindsay’s claim that the operation of models of sufficient complexity to perform real-world tasks cannot be captured with simple statements (2021, p. 2024). Lindsay is probably right. However, we do not think that understanding of phenomena needs to consist in being able to make *simple* statements. Deep ANNs might be challenging to understand. It might be laborious to gain insights about how they work, those insights might not be simple, and establishing their relationship to target phenomena may be difficult. But they might still enable a deeper scientific understanding than highly intuitive but overly simple equations or box-and-arrow models. How easy it is to understand a model does not strictly determine how instrumentally valuable that model will be to understanding and explaining a phenomenon all things considered.

Importantly, partial understanding of a complex but accurate model can lead to greater understanding of a phenomenon than complete understanding of a simple but inaccurate model. To see this, let us focus on prediction for the moment (we will return to the other criteria for understanding in Chapter 5.3). Consider the following toy example: a system *S* elicits a behaviour *P* (the phenomenon of interest). *S* consists of 5 interacting factors (*f*₁, ..., *f*₅) and *P* is described by a single scalar value *y* that can be obtained by simple mathematical operations performed on values taken by *f*₁, ..., *f*₅. Bob constructs a simple model to help him predict *P*. Bob’s model consists of two predictors, *x*₁ and *x*₂, which could (but need not) approximate *f*₁ and *f*₂, respectively. To make predictions about *P*, Bob calculates *x*₁ + *x*₂ in his head, which he can do flawlessly. As

such, we can say that Bob understands his model perfectly; but (due to the simplicity of his model) his predictions about the phenomenon are very inaccurate. Meanwhile, Alice constructs a complex model to help her understand P. Alice's model consists of five predictors $x'_1, \dots, x'_5$ which, with accurate coefficients, precisely capture the factors of the target system: $0.83 \cdot x'_1 + 0.96 \cdot x'_2 + 0.76 \cdot x'_3 + 0.22 \cdot x'_4 + 0.01 \cdot x'_5$. To make predictions about P without running the model (as demanded by the *qualitative* prediction criterion on understanding), Alice estimates $0.8 \cdot x'_1 + x'_2 + 0.8 \cdot x'_3 + 0.2 \cdot x'_4$ in her head. That is, she calculates using imprecise multiplier values and omits x'_5 completely.⁹⁹ Thus, we might say her grasp of the model is partial or incomplete, even though the model itself is perfect. Still, Alice's predictions about P are reasonably accurate. So, despite Alice's model understanding being partial where Bob's is perfect, Alice's predictions about P are still better than Bob's. As such, all else being equal, we attribute greater *model-induced understanding* of P to Alice than to Bob.

This toy example makes it plain that there is an important conceptual distinction between model understanding and model-induced understanding. Even if we can understand a simple model perfectly, we may gain less scientific understanding about a phenomenon P than we can from a complex model of P we only partially understand. Whether we *actually* gain a deeper understanding of P by utilising complex models remains an empirical question. Still, if our diagnosis is correct there is no *principled* trade-off between predictive accuracy and the potential for models to confer scientific understanding. There is no reason to fear the bogeyman! Utilising deep learning in neuroscience need not systematically inhibit our understanding of the brain. Rather, we think, it can support our understanding of neuroscientific phenomena in previously unavailable ways. This takes us to the next section.

⁹⁹ We should think of Alice omitting x'_5 both because it becomes too hard for her to perform a fifth multiplication in her head and because she recognises it will not have a huge effect on the system's behavior. This detail about x'_5 also mirrors cases where parts of the internal structure of ANNs remain opaque.

5.3 Understanding the Brain with Deep Learning

So far, we suggested that scientific understanding consists in having a qualitative grasp of target phenomena and that this involves being able to predict, control, and answer questions about the phenomena in question. Further, we claimed that such understanding is often model-induced. That is, scientists come to understand phenomena through the construction, use, and analysis of models; this applies to simple equations or box-and-arrow-style models as well as to contemporary deep learning-based models. In this section, we elaborate on how ANNs may contribute to an improved understanding of neuroscientific phenomena. First, we argue that deep learning-based models offer a novel *epistemic perspective* which may render accessible previously inaccessible aspects of neuroscientific phenomena (Chapter 5.3.1). Second, we argue that understanding complex phenomena typically requires integrating *distributed explanatory information* produced from multiple epistemic perspectives (Chapter 5.3.2). With this conceptual machinery in place, we sketch a specific research strategy by which ANN-driven neuroscience may contribute to researchers' understanding of biological brains (Chapter 5.3.3).

5.3.1 A novel epistemic perspective

Epistemic perspectives are best described as ways in which researchers approach a target phenomenon – through the use of tools, skills, and theoretical assumptions – to yield explanatory information about how it is elicited by a system. Figuratively speaking, we may think of an epistemic perspective as the outlook or view a scientist takes on a phenomenon; it is constrained by their specific knowledge and capacities, as well as their research questions. More specifically, epistemic perspectives can be characterised with respect to the following dimensions: i) spatiotemporal granularity, ii) specificity, i.e., which kinds of entities and dependencies are detected and described, iii) point of view, including ontological and theoretical assumptions, iv) sensitivity to different factors, and v) scope, i.e., class of phenomena which are investigated (Kästner, 2018, p. 74). Crucially, different epistemic perspectives need not be conceived as competing with each other. Although research from a particular perspective is sometimes sufficient for answering specific questions, deeper understanding of complex systems requires figuring out how

insights from multiple epistemic perspectives fit together (Mitchell, 2002, 2019). Before we consider the question of how multiple perspectives can be integrated (Chapter 5.3.2), let us briefly sketch the core features of the epistemic perspective a deep learning approach to computational neuroscience provides. Needless to say, this will be unavoidably simplistic. In reality, practitioners constructing ANN models of neural functions may endorse various theoretical commitments. Still, it is valuable to see where the contrast lies to more biologically-grounded neuroscience research.

Spatiotemporal Granularity. Traditional neuroscience methods are constrained with respect to both spatial and temporal resolution, though in various ways (Churchland & Sejnowski, 1988; Grinvald & Hildesheim, 2004). Similarly, the deep learning approach to computational neuroscience can be applied at numerous spatial scales, e.g., by targeting neural populations in specific anatomical regions (Sussillo et al., 2012) or distributed processing along pathways like the ventral visual stream (Yamins et al., 2014). What is novel, though, is that *the same model* can shed light on neural phenomena across a wide range of temporal scales. For example, one model can be taken to represent processes extending all the way from a single perceptual judgement (i.e., the inference of the trained model) to evolutionary learning (i.e., the learning trajectory the system undertakes during training) (Cao & Yamins, 2021; Zador, 2019).

Specificity. Being based on automated learning procedures, the deep learning approach has the potential to pick up on patterns in data and dependencies between factors that a human researcher will not usually see or hypothesise. A case in point are functional groups of neurons in V4 we discuss below (Chapter 5.3.3). For a detailed discussion of how deep learning-based models can help uncover structures inaccessible from other epistemic perspectives see Kästner & Crook (manuscript).

Point of View. The deep learning approach casts the brain as an optimisation machine, with specialised modules solving individual computational problems (Cao & Yamins, 2021; Marblestone et al., 2016). This can be contrasted with more biologically-grounded approaches to neuroscience which reject some of these theoretical assumptions (Buzsáki, 2006; Pessoa, 2023; Pulvermüller et al., 2021).

Sensitivity. Unlike biologically-grounded neuroscience, which may treat any and all features of neuroanatomy as relevant to understanding the brain (including, e.g., glial

cells or endocrine signals), the deep learning approach is primarily sensitive to features which are necessary for solving the computational task at hand. Further, because practitioners are interested in how *biological brains* solve these tasks, they especially value models which replicate characteristic error profiles in behavioural tasks (say, object recognition) and display similar patterns of neural activity (as measured by, e.g., fMRI) (Bowers et al., 2022; Schrimpf et al., 2020).

Scope. The flexibility of the deep learning approach renders it amenable to shedding light on a broad range of cognitive and perceptual phenomena, including vision (Zhuang et al., 2021), audition (Kell et al., 2018), language comprehension (Schrimpf et al., 2020) working memory (Kozachkov et al., 2022), and many more besides (see Cichy & Kaiser, 2019; Doerig et al., 2023 for discussion). Unlike biologically-grounded neuroscience, this research is not constrained by ethical concerns or the availability of model organisms. Thus, the deep learning approach has a broader scope allowing practitioners to study, for instance, the “evolutionary goals and historical or developmental constraints that are responsible for shaping a system” (Cao & Yamins, 2021, p. 2; see also Lillicrap & Kording, 2019). These dependencies can be illuminated by focusing on the *designed* components of ANNs – objective functions, architecture, and learning algorithms – which do not suffer from problems of opacity (Richards et al., 2019; Thompson, 2021). In addition, we think deep learning can also help uncover the functional and computational structure of (biological as well as artificial) neural networks (Chapter 5.3.3).

In summary, our exposition makes it plain that the deep learning approach provides a novel epistemic perspective on biological neural processing which renders accessible information unavailable from other epistemic perspectives (i.e., those typically taken by more biologically-grounded neuroscientists). As such, we think it crucially complements existing research practices aiming to understand the brain. To unlock its full potential, however, insights from the deep learning approach must likely be combined with those gained from traditional modelling and biologically-grounded neuroscience.

5.3.2 Distributed explanation, integration, and constraints

We already hinted that the relationship between understanding models and understanding phenomena is complex. One major reason this is so, we take it, is that *in*

practice a subject S's understanding a phenomenon P will not usually rely on just a single model. According to Hochstein, "an individual model is rarely applied in isolation, and is often used to complement a huge body of background information and pre-existing models about the target system" (2016, p. 1401). Recall that for S to understand P involves the abilities to (i) make coarse-grained predictions about P, (ii) give rich, detailed, and productive answers to a broad range of questions concerning P, and (iii) intervene on and control P (see Chapter 5.2.2). To achieve all of this, it will usually be required for S to draw on a *distributed body of information* about P and the system eliciting it.

To clarify, it will be useful to briefly talk about distributed explanations – where the idea of a distributed body of information has been explored. Distributed explanations consist, at least partially, in sets of models, where each model provides different explanatory information about a given phenomenon by illuminating it from different epistemic perspectives (cf., Craver & Kaplan, 2020; Hochstein, 2016; Veit, 2020). If we are correct that understanding of phenomena is routinely supported by distributed explanations, philosophers limiting their analyses to particular models are making a mistake (Veit, 2020, p. 108).¹⁰⁰ Individual models only ever yield partial understanding of complex phenomena. While we agree that sets of models are key constituents of distributed explanations, we see no reason to stop there. The *explanatory store* which facilitates scientific explanation and – as far as we are concerned – understanding should also admit data graphs, diagrams, explanatory writing in natural language, and any other representational vehicles that support understanding (Burnston, 2016; Kästner, 2018; Kohár & Krickel, 2021).¹⁰¹ Thus, the items that support scientific understanding (viz. the elements of the explanatory store) may be of very different kinds and, just as models, may come in varying degrees of complexity.

Despite the looming messiness of the overall situation, we think it is important to acknowledge both the distributed character of the information which supports scientific understanding and the varying contributions that different elements in the explanatory

¹⁰⁰ Naturally, focusing on particular models is only a mistake for cases in which the aims of the analysis *demand* a more holistic view. We think the debate over deep learning's contribution to scientific understanding is such a case.

¹⁰¹ The term *explanatory store* is originally due to Kitcher (1981) and is also used by Craver and Kaplan (2020).

store can make to it. First and foremost, this is because it accords with scientific practice. For instance, Craver and Kaplan observe that “many models are involved, explicitly or implicitly, in most explanations” and that we use various models to “gesture towards explanations that invariably have more content about the causal structure of the world than any single, useful model can express” (2020, pp. 309–310). Besides, once we acknowledge that the information that supports scientific understanding is distributed, we can circumvent undesirable trade-offs inherent to model building (Matthewson & Weisberg, 2009). There is no principled reason we cannot enjoy predictive success *and* improved understanding if we pool insights from multiple models in a common explanatory store.

Unfortunately, these benefits do not come for free. Reaping them requires at least a provisional account of *how* different items in the explanatory store (viz. information generated from multiple epistemic perspectives) can support greater understanding than mastery of particular models (cf., Chang, 2012, Chapter 5). The answer, we think, is that many items in the explanatory store not only contribute explanatory information but also *constraints* (Lawler & Sullivan, 2021; Sullivan, 2022). Based on these constraints, insights provided by different models (or other items) may be integrated with one another (see also Kästner, 2018).

Focusing on models specifically, Mitchell states that each model “provides a partial grasp of the phenomenon, and each requires input and ongoing engagement with the other perspectives” (2019, p. 179). In our view, the engagement Mitchell describes often consists of *constraining interpretation*. The interpretation of each source of explanatory information is constrained by the rest of the explanatory store (or, for cognitive tractability, a small but relevant subset). The exact way this plays out in any single case is subtle, depending crucially on the evidentiary support different sources of explanatory information enjoy, their perceived relevance to the phenomenon at hand, and what aspects of the phenomenon scientists are particularly interested in. For illustration, we briefly present two cases that highlight ways in which the deep learning approach can support scientific understanding of biological neural processing. Though the cases are different, they each involve contributing insights from a novel epistemic perspective and integrating them with other evidence.

Case 1: Vision

Bowers and colleagues' (2022) make the case that the deep learning-based approach to vision neuroscience has paid inadequate attention to research documenting the idiosyncratic features of biological vision. The failure to adapt computational models to these constraints has resulted in the development of models which fail to replicate observed experimental effects in human vision. For instance, while humans show a clear preference for shape over texture, vision ANNs show the opposite pattern (Geirhos et al., 2018). As a result, Bowers and colleagues argue, we cannot (yet) utilise these models to gain insights into the algorithmic structure of biological vision. However, in light of evidence showing similar functional specialisation of single neurons in ANNs and biological visual systems (Pospisil et al., 2018; Willeke et al., 2023), we think a more nuanced conclusion should be drawn. Our understanding of human vision should provisionally integrate *some* properties of vision ANNs (e.g., the hierarchical composition of representations and the functional specialisation of units), while leaving others out (e.g., early layer neurons' overly strong preference for texture). Then, by systematically varying ANN architectures (e.g., by introducing recurrence, Kar et al., 2019) and training procedures (e.g., switching from supervised to unsupervised learning, Zhuang et al., 2021) and comparing the resulting models with diverse sources of neuroscientific data, researchers can clarify which features of ANNs are really i) biologically plausible, and ii) functionally relevant. This way, neuroscientists can refine predictions about how neural populations will respond to certain stimuli and target loci of control (Bashivan et al., 2019). This effectively means integrating research from the epistemic perspectives of traditional vision neuroscience and the deep learning approach to increase our understanding of biological vision.

Case 2: Working memory

Kozachkov and colleagues (2022) use recurrent neural networks (RNNs) to explore the hypothesis that short-term synaptic plasticity (STSP) plays a functional role in working memory (Mongillo et al., 2008). The authors trained RNNs both with and without an STSP mechanism to maintain items in working memory over a delay period. They then assessed them with respect to three criteria: 1) the robustness of their performance in the face of distractors, 2) graceful degradation to synaptic loss, 3) similarity to the neural

dynamics of the prefrontal cortex. Kozachkov and colleagues found that both types of RNN were robust to distractors, but that those with the STSP mechanism degraded more gracefully and exhibited dynamics closer to those of the (primate) prefrontal cortex. In other words, the epistemic perspective researchers adopted by applying the deep learning approach allowed them to reveal previously unknown dependencies between architectural and functional properties of neural structures. Put more generally, we might say that the deep learning approach lent credence to particular computational and mechanistic hypotheses about human brain function (e.g., Miller et al., 2018; Mongillo et al., 2008; see also Doerig et al., 2023). Crucially, in this case, too, it is the integration of insights from different epistemic perspectives that helps researchers advance their understanding of complex phenomena such as working memory.

Thus far, we argued that deep learning-based models offer an additional epistemic perspective for brain researchers that can complement and constrain insights from traditional modelling or biologically-grounded neuroscience by contributing to a distributed store of knowledge. An argument to the same effect could have been made about other innovations in research, e.g. the invention of fMRI scans or the availability of even simple machine learning methods. Against this backdrop readers may wonder: is there anything really *special* about ANN-driven neuroscience?

5.3.3 Understanding biological brains through re-engineering

We already argued that we should not fear the bogeyman, but is there reason to welcome him with open arms? We think there is! By combining the strengths of deep learning with post-hoc interpretability techniques and biologically-grounded neuroscience, researchers can access the patterns embedded in trained ANNs and use them to gain insight into the functional and computational structure of neurobiological systems. Notably, the mechanistic description of biological systems this enables is precisely what critics fear might go missing when researchers utilise ML-based models to try and understand the brain (Chirimuuta, 2021; López-Rubio & Ratti, 2021; Thompson, 2021). We do not think that the deep learning approach sacrifices functional and computational structures; rather, it sheds a new light on them.

To see why this is and how it might work, recall that deep learning models are usually opaque. While the learned parameters of the models are accessible, understanding how they represent information and implement the mapping from input to output requires a dedicated research effort (i.e., this information is – unlike with traditional scientific models – not *given*, cf. Chapter 5.2.1). To uncover the functional structure and computational properties of trained ANNs, the application of systematic post-hoc interpretability techniques is needed (e.g., Bau et al., 2017; Cammarata et al., 2020; Geiger et al., 2021, 2022; Nanda et al., 2023; Olah et al., 2018). As of yet (and as far as we are aware), the research area that most vigorously employs post-hoc interpretability techniques is explainable AI (XAI), not neuroscience (for theoretical discussion of XAI see Chapters 6 and 7 in Durán & Pozzi (Eds.), forthcoming). While we are certainly not the only ones to think that XAI presents opportunities for scientific discovery generally (see Zednik & Boelsen, 2022), we believe that applying post-hoc interpretability techniques to trained ANN models provides a unique opportunity for neuroscience by facilitating an *iterative re-engineering process*.

The process we have in mind involves the following key steps. First, train an ANN to implement an ethologically relevant task (as described in Yamins & DiCarlo, 2016). Second, probe the behavioural characteristics of the trained ANN and evaluate the results with respect to explanatory information distributed across the explanatory store (as in Bowers et al., 2022).¹⁰² Third, employ systematic post-hoc interpretability techniques to uncover the functional and computational structure of the trained ANN (as in Cammarata et al., 2020). Fourth, use these findings to develop hypotheses about the target biological system (Willeke et al., 2023). Fifth, use a variety of more traditional neuroscientific methods to test whether those hypotheses apply to biological target systems. Each step in the process may need to be refined and repeated multiple times in light of the constraints provided by the other steps.

The greatest source of uncertainty over the promise of the iterative re-engineering process is the plausibility of the third step, viz. the usefulness of post-hoc interpretability

¹⁰² Obtaining promising results in this step may be a pre-requisite for engaging in the laborious and challenging process of reverse-engineering an ANN. As such, steps 1 and 2 already form a mini-loop which may be iterated over numerous times before step 3 is reached.

techniques (Cearns et al., 2019). However, recent work by a group around Chris Olah (Cammarata et al., 2020, 2021; Elhage et al., 2021; Olah et al., 2018) serves as a proof of principle that the kind of strategy we have in mind is applicable in practice. The researchers set out to characterise the functional structure of the image classification ANN *InceptionV1* (note that this ANN was not developed explicitly as a model of biological vision). They applied systematic post-hoc interpretability techniques to reverse engineer the opaque system. Focusing on curve detection specifically they found that:

“although curve detection involves more than 50,000 parameters, those parameters actually implement a simple algorithm that can be read off the weights and described in just a few English sentences.” (Cammarata et al., 2021)

Indeed, the detailed understanding of the system's functional structure enabled the researchers to re-design (i.e., hand-code) this part of the system (i.e., a curve circuit) from scratch. Subsequently, they investigated – with research strategies familiar from biologically-grounded neuroscience – to what extent this re-engineered system actually displayed the same properties as the original system (for a detailed discussion see Chapter 3). This way, they validated their interpretable model of inceptionV1's curve detectors. Though this kind of research is still in its infancy, it clearly illustrates that re-engineering these ANNs through the application of post-hoc interpretability techniques *is possible*. As such, we think doing so on ANNs trained on ethologically relevant tasks might provide a fruitful way to uncover functional and structural features that can be converted into hypotheses for brain research.

Indeed, we know of one case in which this research has already inspired hypothesis generation. After synthesising images (MEIs) to maximally excite V4 neurons in the macaque visual system, Willeke and colleagues (2023, p. 2) observed that units were arranged into functional groups selective for various visual features including curves, oriented textural patterns, and eye-like shapes. Referring to Olah's work, these researchers noted a “striking similarity between [the] single cell MEIs of V4 neurons and

single units in the InceptionV1 architecture” (2023, p. 10).¹⁰³ Building on these similarities, Willeke et al. derived novel hypotheses about feature selectivity in the visual processing areas of the primate brain. More specifically, they suggest that ANNs can be used to “derive predictions about colour boundary encoding in monkey V4 functional groups, which could subsequently be verified in in vivo experiments” (2023, p. 7). These descriptions capture steps 3-5 of our iterative re-engineering process precisely.

Naturally, which precise models, tools and theories will be utilised most effectively throughout such a process obviously remains an empirical question. Indeed, it will likely vary from case to case. Still, we believe that we have shown that the iterative re-engineering research strategy has great potential. It may improve scientists’ ability to make coarse qualitative predictions about brain behaviour, answer questions about biological neural processing, and design effective interventions on the brain. That is, it provides an opportunity to foster our understanding of the brain in yet unimagined ways.

5.4 Conclusion

The aim of (neuro)science is to foster understanding of phenomena in the world. Machine learning approaches, particularly deep learning-based ANNs, can contribute to this endeavour by offering a novel and unique epistemic perspective affording insights into the operation of complex systems that are otherwise unavailable. While some have been worrying that this strategy implies trading one black box for another and might actually hinder scientific understanding, we argued these worries are unfounded. There is no trade-off between the predictive success of a model and how much understanding it can confer. Thus, utilising complex computational models in neuroscience will not generally inhibit our ability to understand the (human) brain. Quite to the contrary, we believe, deep learning is best conceived as complementary to established neuroscientific methodology. It offers a powerful addition to neuroscientists’ toolkit enabling constraint-based integration of numerous complementary epistemic perspectives. We thus conclude that deep learning contributes to – rather than detracts from – understanding

¹⁰³ Note that Willeke and colleagues performed psychophysics experiments and quantitative analyses to verify the robustness of their qualitative similarity judgements.

the phenomena neuroscientists investigate. While our argument has focused on the domain of neuroscience specifically, we suggest that it is likely to carry over to relevantly similar scientific domains, even if the details may differ.

5.5 References

- Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science (New York, N.Y.)*, 364(6439), eaav9436. <https://doi.org/10.1126/science.aav9436>
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). *Network Dissection: Quantifying Interpretability of Deep Visual Representations*. 3319–3327. <https://doi.org/10.1109/CVPR.2017.354>
- Bechtel, W., & Richardson, R. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. <https://doi.org/10.7551/mitpress/8328.001.0001>
- Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The Two Cultures of Computational Psychiatry. *JAMA Psychiatry*, 76(6), 563–564. <https://doi.org/10.1001/jamapsychiatry.2019.0231>
- Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1), 43–75. <https://doi.org/10.1007/s11023-021-09569-4>
- Bokulich, A. (2017). Models and Explanation. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science* (pp. 103–118). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4_4
- Boon, M. (2020). How Scientists Are Brought Back into Science—The Error of Empiricism. *A Critical Reflection on Automated Science: Will Science Remain Human?*, 43–65. https://doi.org/10.1007/978-3-030-25001-0_4
- Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section D, Structural Biology*, 77(Pt 8), 982–991. <https://doi.org/10.1107/S2059798321007531>
- Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolphi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *The Behavioral and Brain Sciences*, 1–74. <https://doi.org/10.1017/S0140525X22002813>
- Burnston, D. C. (2016). Data graphs and mechanistic explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 57, 1–12. <https://doi.org/10.1016/j.shpsc.2016.01.002>
- Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195301069.001.0001>
- Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, 5(3), 10.23915/distill.00024. <https://doi.org/10.23915/distill.00024>

- Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve Circuits. *Distill*, 6(1), e00024.006. <https://doi.org/10.23915/distill.00024.006>
- Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. *ArXiv:2104.01489 [Cs, q-Bio]*. <http://arxiv.org/abs/2104.01489>
- Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, 9(1), Article 1. <https://doi.org/10.1038/s41398-019-0607-2>
- Chang, H. (2012). *Is Water H₂O? Evidence, Realism and Pluralism*. Boston Studies in the Philosophy and History of Science.
- Charles Leek, E., Leonardis, A., & Heinke, D. (2022). Deep neural networks and image classification in biological vision. *Vision Research*, 197, 108058. <https://doi.org/10.1016/j.visres.2022.108058>
- Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199(1), 767–790. <https://doi.org/10.1007/s11229-020-02713-0>
- Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Simon and Schuster.
- Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science (New York, N.Y.)*, 242(4879), 741–745. <https://doi.org/10.1126/science.3055294>
- Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4), 305–317. <https://doi.org/10.1016/j.tics.2019.01.009>
- Colaco, D. (2019). *An Investigation of Scientific Phenomena* [PhD Thesis]. University of Pittsburgh.
- Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, 68(1), 53–74.
- Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.
- Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press. <https://press.uchicago.edu/ucp/books/book/chicago/l/bo16123713.html>
- Craver, C. F., & Kaplan, D. M. (2020). Are More Details Better? On the Norms of Completeness for Mechanistic Explanations. *The British Journal for the Philosophy of Science*, 71(1), 287–319. <https://doi.org/10.1093/bjps/axy015>
- Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4), 568–589. <https://doi.org/10.1086/709729>
- Crook, B., Schlüter, M., & Speith, T. Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)* (pp. 316-324). IEEE.
- de Regt, H. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, 76(5), 585–597. <https://doi.org/10.1086/605795>
- Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1–20. <https://doi.org/10.1038/s41583-023-00705-w>

- Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, 132(1), 33–42.
<https://doi.org/10.1007/s11098-006-9054-z>
- Elgin, C. (2017). *True enough*. MIT Press.
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., ... Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.
- Feest, U. (2017). Phenomena and Objects of Research in the Cognitive and Behavioral Sciences. *Philosophy of Science*, 84(5), 1165–1176.
- Frigg, R., & Hartmann, S. (2020). Models in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2020). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>
- Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, 34, 9574–9586.
<https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html>
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing Causal Structure for Interpretable Neural Networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338.
<https://proceedings.mlr.press/v162/geiger22a.html>
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018, December 21). *ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness*. International Conference on Learning Representations. <https://openreview.net/forum?id=Bygh9j09KX>
- Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, 175, 126–137.
<https://doi.org/10.1016/j.pneurobio.2019.01.008>
- Glennan, S. (2017). *The new mechanical philosophy* (First edition). Oxford University Press.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Grimm, S. (2021). Understanding. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University.
<https://plato.stanford.edu/archives/sum2021/entries/understanding/>
- Grinvald, A., & Hildesheim, R. (2004). VSDI: A new era in functional imaging of cortical dynamics. *Nature Reviews. Neuroscience*, 5(11), 874–885.
<https://doi.org/10.1038/nrn1536>
- Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2), Article 2. <https://doi.org/10.1609/aimag.v40i2.2850>
- Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2), 245–258.
<https://doi.org/10.1016/j.neuron.2017.06.011>
- Hills, A. (2016). Understanding Why. *Noûs*, 50(4), 661–688.
<https://doi.org/10.1111/nous.12092>

- Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, 193(5), 1387–1407. <https://doi.org/10.1007/s11229-015-0844-8>
- Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626. <https://doi.org/10.1007/s11229-008-9435-2>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), Article 7873. <https://doi.org/10.1038/s41586-021-03819-2>
- Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask ‘why’ questions of minds and brains. *Trends in Neurosciences*, 46(3), 240–254. <https://doi.org/10.1016/j.tins.2022.12.008>
- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. *Nature Neuroscience*, 22(6), Article 6. <https://doi.org/10.1038/s41593-019-0392-5>
- Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, 68, 68–79. <https://doi.org/10.1016/j.shpsa.2018.01.011>
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>
- Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science*, 48(4), 507–531.
- Kohár, M., & Krickel, B. (2021). Compare and Contrast: How to Assess the Completeness of Mechanistic Explanation. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience* (pp. 395–424). Springer International Publishing. https://doi.org/10.1007/978-3-030-54092-0_17
- Kozachkov, L., Tauber, J., Lundqvist, M., Brincat, S. L., Slotine, J.-J., & Miller, E. K. (2022). Robust and brain-like working memory through short-term synaptic plasticity. *PLOS Computational Biology*, 18(12), e1010776. <https://doi.org/10.1371/journal.pcbi.1010776>
- Lawler, I., & Sullivan, E. (2021). Model Explanation Versus Model-Induced Explanation. *Foundations of Science*, 26(4), 1049–1074. <https://doi.org/10.1007/s10699-020-09649-1>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. <https://doi.org/10.1038/nature14539>
- Lillicrap, T. P., & Kording, K. P. (2019). *What does it mean to understand a neural network?* (arXiv:1907.06374). arXiv. <https://doi.org/10.48550/arXiv.1907.06374>
- Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

- López-Rubio, E., & Ratti, E. (2021). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, 198(4), 3131–3156. <https://doi.org/10.1007/s11229-019-02271-0>
- Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023, September). Sources of opacity in computer systems: Towards a comprehensive taxonomy. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)* (pp. 337–342). IEEE.
- Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, 10. <https://www.frontiersin.org/article/10.3389/fncom.2016.00094>
- Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, 170(1), 169–190. <https://doi.org/10.1007/s11229-008-9366-y>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/BF02478259>
- Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, 100(2), 463–475. <https://doi.org/10.1016/j.neuron.2018.09.023>
- Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy*, 17(1), 55–70. <https://doi.org/10.1023/A:1012990030867>
- Mitchell, S. D. (2019). Perspectives, Representation, and Integration. In *Understanding Perspectivism* (pp. 178–193). Routledge. <https://doi.org/10.4324/9781315145198-11>
- Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, 319(5869), 1543–1546. <https://doi.org/10.1126/science.1150769>
- Morgan, M. S., & Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.
- Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). *Progress measures for grokking via mechanistic interpretability* (arXiv:2301.05217). arXiv. <https://doi.org/10.48550/arXiv.2301.05217>
- Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010. <https://doi.org/10.23915/distill.00010>
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2. ed). Cambridge Univ. Press.
- Pessoa, L. (2023). The Entangled Brain. *Journal of Cognitive Neuroscience*, 35(3), 349–360. https://doi.org/10.1162/jocn_a_01908
- Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). ‘Artiphsiology’ reveals V4-like shape tuning in a deep network trained for image classification. *ELife*, 7, e38242. <https://doi.org/10.7554/eLife.38242>
- Potochnik, A. (2016). *Scientific Explanation: Putting Communication First*. 721–732.
- Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., & Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22(8), Article 8. <https://doi.org/10.1038/s41583-021-00473-5>

- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
- Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700, pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1
- Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3), 413–423. <https://doi.org/10.1016/j.neuron.2020.07.040>
- Shagrir, O., & Bechtel, W. (2017). Marr's Computational Level and Delineating Phenomena. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780199685509.003.0009>
- Srećković, S., Berber, A., & Filipović, N. (2022). The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. *Minds and Machines*, 32(1), 159–183. <https://doi.org/10.1007/s11023-021-09575-6>
- Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, 44(3), 510–515. <https://doi.org/10.1016/j.shpsa.2012.12.005>
- Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1), 109–133. <https://doi.org/10.1093/bjps/axz035>
- Sussillo, D., Nuyujukian, P., Fan, J. M., Kao, J. C., Stavisky, S. D., Ryu, S., & Shenoy, K. (2012). A recurrent neural network for closed-loop intracortical brain-machine interface decoders. *Journal of Neural Engineering*, 9(2), 026027. <https://doi.org/10.1088/1741-2560/9/2/026027>
- Thompson, J. A. F. (2021). Forms of explanation and understanding for neuroscience and artificial intelligence. *Journal of Neurophysiology*, 126(6), 1860–1874. <https://doi.org/10.1152/jn.00195.2021>
- Veit, W. (2020). Model Pluralism. *Philosophy of the Social Sciences*, 50(2), 91–114. <https://doi.org/10.1177/0048393119894897>
- Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., Cadena, S. A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., Sinz, F. H., & Tolia, A. S. (2023). *Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization* (p. 2023.05.12.540591). bioRxiv. <https://doi.org/10.1101/2023.05.12.540591>
- Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.
<https://doi.org/10.1038/nn.4244>

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.
<https://doi.org/10.1073/pnas.1403112111>

Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1), 3770.
<https://doi.org/10.1038/s41467-019-11786-6>

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288.
<https://doi.org/10.1007/s13347-019-00382-7>

Zednik, C., & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines*, 32(1), 219–239. <https://doi.org/10.1007/s11023-021-09583-6>

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), e2014196118.
<https://doi.org/10.1073/pnas.2014196118>

Chapter 6:

Understanding as a Bottleneck for the Data-Driven

Approach to Psychiatric Science

This chapter was published as “Crook, B. (2023). Understanding as a bottleneck for the data-driven approach to psychiatric science. *Philosophy and the Mind Sciences*, 4.” Slight changes have been made to fit the dissertation format.

6.0 Abstract

The data-driven approach to psychiatric science leverages large volumes of patient data to construct machine learning models with the goal of optimising clinical decision making. Advocates claim that this methodology is well-placed to deliver transformative improvements to psychiatric science. I argue that talk of a data-driven revolution in psychiatry is premature. Transformative improvements, cashed out in terms of better patient outcomes, cannot be achieved without addressing *patient understanding*. That is, how patients understand their own mental illnesses. I conceptualise understanding as the possession of adaptive mental constructs through which experience is mediated. I suggest that this notion of understanding serves as a bottleneck which any prospective approach to psychiatry must address to be efficacious. Subsequently I argue that, though the data-driven approach is undoubtedly powerful, it does not have a straightforward means of unblocking the bottleneck of understanding. I suggest that the data-driven approach must be supplemented with significant theoretical progress if it is to transform psychiatry.

Heavenly Hurt, it gives us –

We can find no scar,

But internal difference –

Where the Meanings, are –

Emily Dickinson (Dickinson, 1890/1998), *There's a certain Slant of Light*

6.1 Introduction

Mental illness is one of the greatest sources of human suffering worldwide (Insel et al., 2015; Vigo et al., 2016). Governments, research institutions, and health organisations are leading calls for improved mental healthcare to ameliorate the devastating costs that mental illness imposes on individuals and societies (Department of Health and Social Care, 2021; Well Being Trust, 2020; WHO, 2019). Despite frequent claims that current approaches to tackling mental illness are inadequate, theorists are divided about how progress should be made (Bickman, 2020; Cuthbert & Insel, 2013; Hengartner & Lehmann, 2017). Recently, methods utilising big data and artificial intelligence (AI) have been gaining increased attention. Proponents of this *data-driven approach* are hopeful that sophisticated machine learning models will “redesign the current landscape of mental illness” (Fernandes et al., 2017, p. 5) through “tailored interventions for better outcomes” (Fernandes et al., 2017, p. 3). In this paper I critically assess the claim that the data-driven approach, as currently conceived, can bring about a transformative improvement to psychiatric science. I argue that, due to the central role *patient understanding* plays in recovery from mental illness, talk of an AI revolution in psychiatry is premature.

My work follows other authors who have engaged with the philosophical issues raised by the idea of AI-driven psychiatry. In an article structured as a back-and-forth debate, Brown and colleagues (2021) discuss the possibility of future AI replacing human psychiatrists altogether. In a paper focusing on schizophrenia, Starke and colleagues (2021) discuss the ethical principles that must be upheld if machine learning is to be introduced into clinical practice. In a response to that article, Gauld and colleagues (2021) analyse the kind of change that the age of AI will presage for psychiatry and the training of psychiatrists, a topic also addressed by McCoy and colleagues (2020). Finally, Horn and Weisz (2020) tackle the question of whether AI can improve psychotherapy research. My argument covers similar ground to all of the above work. The novelty of my approach consists in the key role it affords patient understanding and how thoroughly it

engages with the evidence on what makes interventions effective. Together, these theoretical and empirical factors have important implications for how successful we can expect the data-driven approach to be.

Many practitioners and theorists worry that psychiatry is a science in crisis, with traditional frameworks insufficient to address the complexity of mental illness (Bickman, 2020; McGorry & Nelson, 2019). Advocates of the data-driven approach are optimistic that their framework is the first to be capable of tackling that complexity in a scientifically rigorous way (Durstewitz et al., 2019; Dwyer et al., 2018; Rutledge et al., 2019). I begin the paper by describing this approach. Methodologically, the data-driven approach eschews traditional hypothesis-driven inferential statistics in favour of the automated discovery of predictive patterns in large volumes of patient data. Philosophically, I note that advocates of this approach espouse *theory neutrality*, rejecting specific theories of mental illness on the grounds that data are a more reliable and objective guide to clinical decision-making. I also characterise the data-driven approach as *function-oriented*, viewing psychiatry as an interdependent set of optimisation problems. I argue that these commitments lead to problems for the approach once the role of patient understanding is carefully considered.

To set up my argument, I begin by clarifying what would constitute a *transformative* improvement to the status quo. I suggest that a reasonable interpretation should focus on patient outcomes, as “the amelioration of ... mental distress and suffering” (Thornton, 2020, p. 237) is the fundamental goal of psychiatry.¹⁰⁴ Given such a view, I argue that for the data-driven approach to be transformative, it must bring about radical improvements in patient outcomes. My argument for tempering expectations of the data-driven approach centers on *patient understanding*. Integrating ideas about understanding from pragmatist philosophers and theorists of psychiatry (Beck & Haigh, 2014; Dilthey, 1894/1977; Wampold & Imel, 2015; Wilkenfeld, 2013), I conceptualise patient understanding as a set of *adaptive mental constructs* through which a subject’s experiences are mediated. Conceptualised this way, a patient possessing understanding

¹⁰⁴ To be clear then, radically improved *scientific understanding* of mental disorders, if unaccompanied by improved outcomes, would not be considered transformative by this criterion. However, one could reasonably adopt another criterion on which it would. I thank an anonymous reviewer for pushing me to clarify this point.

becomes a pre-requisite for lasting recovery from mental disorder. I suggest that this notion of understanding serves as a *bottleneck* for any approach to psychiatric science. Without addressing patient understanding, outcomes cannot be radically improved.

The above analysis leads us to a specific question: can the data-driven approach address the bottleneck of understanding? I examine three ways that it might. First, I assess what I call the *indirect path*, which consists of optimisation of the clinical pipeline in terms of nosology, diagnosis, prognosis, and treatment assignment. Second, I assess what I call the *direct path*. That is, the tailoring of treatments for improved outcomes. Third, I assess whether the data-driven approach might deliver radically improved patient outcomes by discovering novel insights about mental disorders that can be used to provide adaptive explanations. I argue that there are serious problems with each of these possibilities and, therefore, that the data-driven approach has no clear route to providing transformative improvements to psychiatric science.

The structure of the paper is as follows. In Chapter 6.2 I present the data-driven approach. I discuss what motivates it, how it works, and its philosophical commitments. In Chapter 6.3 I make the claim that patient understanding functions as a bottleneck for any approach to psychiatric science. To do so, I present my analysis of understanding as possession of adaptive mental constructs, show how this makes understanding critical for robust recovery from mental distress, and explain in what sense this makes it a bottleneck. In Chapter 6.4 I assess the potential of the data-driven approach to address the challenge of the bottleneck of understanding, arguing that, at least as currently practiced, it is ill-suited to do so. In Chapter 6.5 I address possible objections to the view presented. In particular, I defend my conceptualisation of understanding and justify the applicability of the bottleneck metaphor. I then conclude by suggesting that the data-driven approach can best contribute to psychiatric science in collaboration with theoretical research.

6.2 The Data-driven Approach

In this section I characterise the data-driven approach to psychiatry. I begin by describing the theoretical landscape from which the approach has emerged. I then provide an

overview of the data-driven methodology, including an example (Koutsouleris et al., 2021). I conclude the section by discussing the philosophical commitments embodied by the approach.

6.2.1 Motivating the approach: Psychiatry in crisis

Reading the theoretical and philosophical literature on mental illness, one often comes across the view that psychiatry is a discipline in crisis. For example, Stoyanov and Maes (2021, p. 1) claim that “psychiatry remains in a permanent state of crisis... evidenced by the many different competing approaches and ways to understand mental and psychiatric disorders”. Leonard Bickman (2020, p. 803) states that mental healthcare currently suffers from “overall poor effectiveness” caused by “poor alignment between what the patient needs, and the treatment provided”. And Allen Frances (2009, p. 391) writes that “our understanding of psychopathology is fairly primitive” and “lacks the fundamental understanding of pathogenesis”. These pessimistic assessments locate the precise source of the problem differently, but they all share the opinion that the *complexity* of mental disorders plays a significant role in thwarting the success of psychiatric science.¹⁰⁵

Mental disorders are complex in multiple ways. First, they are difficult to explain, resisting “a simple analysis in terms of biological pathways, endophenotypes, and neural mechanisms” (Borsboom et al., 2019, p. 91). Second, mental disorders are massively multifactorial (de Boer et al., 2021). That is, there are many different factors, “sprinkled across multiple causal levels” (Kendler & Gyngell, 2020, p. 44), that may be relevant to any given presentation of a mental disorder. This relates to the first point. If we ought, as Engel suggested, to incorporate “all the factors contributing to both illness and patienthood” (1977, p. 133) into our explanations of mental disorder, then these will be dense explanations indeed. However, identifying the relevant factors does not suffice to fully characterise mental disorders. This is due to a third kind of complexity: interaction among the factors. As Borsboom and colleagues put it, “mental disorders likely involve feedback loops that cross all of the traditional divides between levels of explanation,

¹⁰⁵ The term complexity is used to refer to many different properties, both formal and informal (Ladyman et al., 2013). It is beyond the scope of this paper to weigh in on which formal criteria for complexity mental disorders satisfy.

none of which can claim the status of ‘basis’ for the others” (2019, pp. 10–11). In summary, biological entities, psychological constructs, and social factors interact in intricate and diverse ways to produce particular instantiations of psychopathology.

The idea that psychiatry is in a crisis of complexity provides a strong rationale for adopting the data-driven approach. For example, Rutledge and colleagues (2019, p. 152) state that “big data and machine learning are uniquely placed to address [the complexities of psychiatric disorders]”. Similarly, British psychiatrists Brown and Story (2021, p. 131) argue that modern technology “enables the capturing of rich, longitudinal, multimodal data, the analysis of which promises vastly improved characterisation of illnesses and their trajectories”. While historical approaches to psychiatry have often been stuck either focusing on a small subset of factors or falling short of scientific standards of evidence (Ghaemi, 2007), the data-driven approach leverages the availability of data and computation to integrate diverse sources of information in a rigorous way. Unlike traditional statistical methods that focus on detecting group differences, machine learning methods are designed to exploit large volumes of multivariate data to make individual predictions (Bennett et al., 2019; Gillan & Whelan, 2017). This means that clinical decisions, such as treatment recommendations, can be tailored to sophisticated statistical representations of individual patients. For this reason, the data-driven approach is sometimes termed *precision psychiatry* (Bzdok & Meyer-Lindenberg, 2018; Fernandes et al., 2017).

Some proponents of the data-driven approach have been forthright in claiming that this approach will suffice to bring about a *transformative* improvement to mental healthcare. For example, Gauld and colleagues (2021, p. 2519) proclaim that “AI will undeniably transform the future of psychiatry”. Similarly, Bickman (2020, p. 803) extols the “revolutionary possibilities of artificial intelligence for improving mental healthcare”. And Fernandes and colleagues (2017, p. 1) state that the data-driven approach “promises to be even more transformative than in other fields of medicine”. This view is grounded in the notion that the data-driven approach will provide models that allow clinicians “to identify the right treatment for each patient, first time around” (Gillan & Whelan, 2017, p. 34). As stated in the introduction, I interpret a transformative improvement to require

radically improved patient outcomes. In order to assess whether the data-driven approach can really achieve this, we will need to take a closer look at how it works.

6.2.2 The methodology of the data-driven approach

I thus turn to the methodology of the data-driven approach. Readers familiar with big-data and machine learning can skip to Chapter 6.2.3 where I will discuss the philosophical commitments of the approach.

The data-driven approach is defined by a methodological hallmark: *the employment of automated computational procedures for discovering predictive patterns in large volumes of data*. Such approaches have become increasingly attractive as the cost of collecting, storing, and manipulating data has decreased (Chekroud et al., 2021). At a very high level, the methodology of the data-driven approach can be described as consisting of four steps. First, a problem is defined. Second, data are acquired. Third, a model is designed and trained. Fourth, the model is deployed. We will take a look at each of these steps in turn.

Problem definition

The first step in the data-driven approach is to define a problem. Machine learning is a technique well-suited to classification, prediction, and clustering problems. In classification, the task is choosing which of a pre-defined set of classes a particular data instance belongs to. In prediction, the task is estimating the future value of a chosen variable. In clustering, the task is organising a dataset into self-similar groups. Models developed according to the data-driven approach have many possible applications in psychiatry. These can be coarsely divided into four (Bzdok & Meyer-Lindenberg, 2018; Dwyer et al., 2018). First, classification models can be used for *diagnosis*. For example, a model might be constructed to determine whether subjects should be diagnosed with major depressive disorder (e.g., Sharma & Verbeke, 2020). However, applying machine learning to diagnosis need not be limited to a binary decision about a single condition. For example, in multiclass logistic regression a model will return a probability distribution over multiple possible diagnoses (Elujide et al., 2021; Qureshi et al., 2016). This makes the data-driven approach amenable to transdiagnostic studies (Pelin et al., 2021).

Second, models employing clustering algorithms can be used to aid *nosology*, the organisation of mental illnesses into a set of categories (Zachar & Kendler, 2017). Within this paradigm, the objective of the model is to find a partitioning of the data that best satisfies a chosen set of statistical criteria. When models are used to find clusters *within* extant disorder categories, this is called sub-typing (Feczko et al., 2019). A third application of the data-driven approach is *prognosis*, the prediction of how an individual's condition will evolve over time. Machine learning models can be used to predict clinically relevant events, such as symptom onset (Koutsouleris et al., 2021). Recurrent neural networks, an architecture particularly well-suited to exploiting temporal regularities, can be used to compute disease trajectories (Shickel et al., 2018; Suhara et al., 2017). Finally, perhaps the most important target application of all is *treatment assignment*, using patient data to decide which therapeutic approach to pursue.¹⁰⁶ Studies have attempted to apply machine learning to predict responses to both pharmacological and psychotherapeutic treatments (Chekroud et al., 2021; Su et al., 2020).

Data acquisition

Machine learning models are typically *data hungry*, meaning that they only learn patterns that are robust and generalisable if supplied with large volumes of data. This makes data acquisition a crucial part of the approach (Dwyer et al., 2018). The reason for these onerous data requirements is known as the *curse of dimensionality*, a term referring to the fact that, as more dimensions of variation are considered, the numerosity of the data required to maintain model performance increases exponentially (Keogh & Mueen, 2017). This means that collecting the quantities of data required to train massively multivariate models of mental illness requires large-scale coordination (Bzdok & Meyer-Lindenberg, 2018).

Since variation at almost every conceivable level of organisation is thought to be relevant to mental illness (Kendler & Gyngell, 2020), data can be collected in innumerable ways,

¹⁰⁶ Initial research in this domain often involves predicting *response* to an individual treatment as opposed to assignment among treatments. Clearly, if reliable estimates of treatment response could be computed across many treatments, this would lead to a straightforward procedure for assigning treatment. Namely, predicting the response to multiple treatments and selecting the best one [indeed, the problems are treated as synonymous in (Chekroud et al., 2021)].

from questionnaires and cognitive tasks, to biomarkers and neuroimaging, to social media and wearable electronics (Balaskas et al., 2021; García-Gutiérrez et al., 2020). The challenges involved vary across modalities. For example, collecting and analyzing neuroimaging data is time-consuming and costly (Najafpour et al., 2021). On the other hand, clinical notes in Electronic Health Record (EHR) data are abundant and accessible, but unstructured (Shickel et al., 2018). Pre-processing diverse data modalities such that they can be integrated by a single model is a technical challenge. Additionally, it is often unclear whether different operationalisations of psychological concepts actually measure the same latent construct (Poldrack & Yarkoni, 2016). Given these difficulties, Schickel and colleagues (2018, p. 24) describe developing a unified representation of such disparate sources of information as the “holy grail of clinical deep learning research”.

Model design and training

Given a rich and voluminous dataset, the next step is to design a model. In the machine learning paradigm this consists of specifying the objective function, architecture, and learning rule of the model. The objective may be to predict remission rates, responsiveness to specific medications, or symptom response trajectories (Bzdok & Meyer-Lindenberg, 2018; Chekroud et al., 2021; Rutledge et al., 2019). Translating the natural language expression of the goal into a mathematically specified objective function is a crucial part of the design phase. The architecture of a model consists of its components and how they are organised. For example, a deep neural network consists of multiple layers of neuron-like computational units equipped with activation functions (Durstewitz et al., 2019), while a decision tree consists of hierarchically organised decision nodes (Feczko et al., 2019). Machine learning researchers have developed a veritable zoo of model architectures with a variety of mathematical and representational properties (Bronstein et al., 2021). In principle, architectures can be designed to exploit the kinds of regularities in the sampled data domain. In the case of high-dimensional, multimodal patient data, however, design principles guiding architectural choices are not yet well established (Si et al., 2021), reflecting the lack of scientific consensus on the nature of mental illness.

Once an objective and architecture are chosen, training takes place. A machine learning model is usually initialised with random parameter values (Chollet, 2021, p. 46). What the model outputs is a function of the input and these parameter values. In a deep neural network, for example, the parameter values are the connection weights between the neurons and each individual neuron's bias. The goal of learning is to find a set of parameter values that optimises the objective. For example, if the objective is to diagnose depression, the model might take neuroimaging and physiological data associated with an individual and output a binary classification reflecting the presence or absence of a diagnosis (Arbabshirani et al., 2017; Sharma & Verbeke, 2020). The model's output is then compared with the true label for the data point. In this case, the label would be provided by a clinician's diagnostic judgement about the patient. Having access to the ground truth label makes this an example of *supervised learning*. Should the machine's output be incorrect, an error signal, often called the *loss*, is computed. This error signal, in conjunction with a learning rule, like stochastic gradient descent, is used to update the parameter values.¹⁰⁷ After many iterations of this process, a model converges on a stable set of parameter values that encode its best approximation to the true function relating the input data to the output labels. In our example, the parameters can be considered as encoding a hypothesised relationship between biological variables and depression. Crucially, to avoid overfitting on the particularities of the data sample used for training, models must always be validated on novel datasets (Bzdok & Meyer-Lindenberg, 2018; Rutledge et al., 2019).

The simple example above is the tip of the iceberg when it comes to applications of the data-driven approach. While a thorough review is out of scope for this paper, it is worth noting the broad categories of *unsupervised* and *semi-supervised* methods, which do not rely on labels (Bengio et al., 2013; Bronstein et al., 2021; Domingos, 2012). In our depression diagnosis case, ground truth labels were provided by human judgement. In addition to being costly and time-consuming to generate, the use of human judgment in producing labels limits the potential value data-driven models can provide (Bickman, 2020). Durstewitz and colleagues (2019) point out that the value of data-driven diagnosis

¹⁰⁷ In practice, one iteration of a training loop usually involves a batch of data instances, rather than just one.

is upper-bounded by how well extant diagnostic labels capture the underlying nature of mental illnesses. For this reason, the data-driven approach “needs to go beyond the mere prediction of symptoms by current diagnostic schemes, but rather has to help refining our diagnoses” (Durstewitz et al., 2019, p. 1591). In line with this idea, researchers have employed unsupervised machine learning methods to discover novel clusters of patients that can be used to predict outcomes, stratify treatments, and increase understanding (Gould et al., 2014; Miranda et al., 2021; Pelin et al., 2021). The crucial point is that, while supervised learning is limited by human experts’ ability to provide useful labels, the data-driven approach as a whole need not be.

Validation and deployment

Once data have been collected and a model has been trained, the next steps are validation and deployment. As of yet, the majority of studies implementing the data-driven approach are proof-of-concept and most trained models have not been clinically deployed (Cearns et al., 2019; Chekroud et al., 2021). In an insightful review, Cearns and colleagues (2019) provide a checklist that any machine learning model should comply with to demonstrate readiness for clinical application. Models should show: 1) an improvement on whatever approach constitutes the current state-of-the-art for the clinical application in question, 2) validation with a large, external dataset, and 3) clear specification of the scope of the model (i.e., the population for which its validity has been rigorously demonstrated). Once deployed, a machine learning model must be continually evaluated to ensure that its accuracy remains robust, that it is secure, and that it does not lead to unacceptable algorithmic bias. Practically, deployment also requires coordination with clinicians to ensure that models are well-utilised (Tonekaboni et al., 2019). As the data-driven approach matures, further concrete issues relating to clinical deployment are sure to emerge.

The data-driven approach in practice: An example

An example will help to elucidate how the data-driven approach plays out in practice. In a large-scale study, Koutsouleris and colleagues (2021) used multimodal machine learning to predict whether patients with clinical high-risk states and recent-onset depression would transition to psychosis. In an extensive data acquisition process spanning five countries and more than three years, Koutsouleris and colleagues

collected data from many domains including sociodemographic, questionnaires, self-reports, structured interviews, cognitive and behavioural tasks, neuroimaging, and genetics. This highlights the importance of collecting spatially and temporally distributed data for constructing generalisable models that will be robust to variation in social and cultural factors (Bzdok & Meyer-Lindenberg, 2018; Dwyer et al., 2018). To make use of their avalanche of information, Koutsouleris and colleagues built a *stacked* model, a kind of architecture suitable for automatically learning how to best integrate disparate data sources.¹⁰⁸ The researchers' multimodal predictive model was able to outperform all unimodal competitors as well as human clinical raters in prognostic accuracy. In addition to the predictive value of their model, the authors were also able to extract potential insights into psychosis by analyzing their findings. For example, they found high-risk patients predicted *not* to transition to psychosis had increased temporo-occipital brain volume relative to healthy controls, suggesting a neural basis for mechanisms of resilience. Further, they found certain neurocognitive factors, such as impaired facial affect recognition, were markers of poor psychosis outcomes in both clinical high-risk and recent-onset depression groups. This highlights another important feature of the data-driven approach. Investigating a transdiagnostic population, that is, incorporating data from groups which are typically investigated separately, allows the data-driven approach to find and exploit patterns which transcend the group differences captured by traditional inferential approaches (Hengartner & Lehmann, 2017; McGorry & Nelson, 2019).

6.2.3 The philosophical commitments of the data-driven approach

Now that we have covered the methodological basis of the data-driven approach, we can turn to its philosophy. I raise two points. First, the data-driven approach purports to be *theory neutral*, relying on automated statistical procedures rather than theoretical assumptions to cut through the complexity of mental illness. Second, the data-driven approach is *function-oriented*, focusing on building models to optimise each step of the clinical decision-making pipeline. These features will turn out to be important for

¹⁰⁸ Stacked models consist of multiple *base* models, each exploiting a particular data modality, and a *meta-model* which learns how to weight the predictions of the base models.

assessing whether the approach can transform psychiatry by radically improving patient outcomes.

Theory neutrality

Different schools of thought have long disagreed about the *nature* of mental illnesses (Dembić, 2023; Kendler, 2016; Leder & Zawidzki, 2023; Shorter, 1997; Tsou, 2021). Should we think of them as brain disorders defined by neural dysfunction, socially constructed labels for deviant behaviour, irreducibly subjective pathologies of psychology, or an intricate mixture of these ingredients? The data-driven approach chooses to wash its hands of the issue entirely, preferring to remain *theory neutral* (Paulus, 2015). Unlike the attempt to develop mechanistic models of disease processes, the data-driven approach to psychiatry makes no strong assumptions about the nature or causal structure of mental illness (Bennett et al., 2019). Practitioners take this feature to be a strength of the approach. For example, Chekroud and colleagues (2021, p. 154) claim that machine learning offers a set of “powerful hypothesis-free approaches” to predicting treatment outcomes. Similarly, in a review of techniques applying deep learning to Electronic Health Record data, Shickel and colleagues (2018, p. 17) describe the philosophy as “letting the data speak for itself by discovering latent relationships and hierarchical concepts from the raw data, without any human supervision or prior bias”. From the data-driven perspective, then, theoretical assumptions about mental illness do more harm than good, and we would be better off relying on the presumed objectivity of models trained with automated procedures.

It is important to note that the data-driven approach is not, and indeed no approach can be, truly atheoretical (Joober & Tabbane, 2019). In practice, even huge quantities of data are useless without significant constraints on learning (Domingos, 2012). Researchers developing machine learning models make many choices about which data to collect, which architecture to use, how to specify their objective, and so on (see Chapter 6.2.2). Each of these decisions biases a model towards learning some structures rather than others, a phenomenon often discussed in machine learning under the term *inductive bias* (Battaglia et al., 2018). Despite this caveat, the intention to remain as neutral as possible with respect to theory still constitutes a significant philosophical commitment of the data-driven approach. In the argument to come, I will suggest that this is a mistake.

Without theory-driven research to improve understanding of how treatments can improve outcomes, the data-driven approach has limited potential.

Function orientation

Powerful methodological tools alter how domains of inquiry are perceived by their users. Cichy and Kaiser (2019, p. 312) express this point thusly, “once technological artefacts are commonly used, they are not mere tools to realise predefined scientific goals but begin to shape social reality in a way that affects the user’s desires and interests”. The data-driven approach is no exception. As Dwyer and colleagues put it (2018, p. 95), “modern machine learning methods can contribute greatly to clinical psychology and psychiatry *by changing the way that problems are considered* [emphasis added]”. I introduce the term *function orientation* to describe the perspective on psychiatry that the data-driven approach imposes. The term *function* here should be interpreted in the mathematical sense of mappings from inputs to outputs. Note that the function-oriented perspective is largely implicit in the way data-driven research is carried out, rather than being explicitly endorsed by all of its practitioners.

The function-oriented perspective decomposes psychiatric science into a set of statistical problems, each addressable through the development of the right machine learning model. It formulates these problems in mathematical terms, with solutions consisting of trained models judged by accuracy metrics. The overall goal is to optimise the entire pipeline of clinical decision making, from diagnosis to treatment assignment. As Dwyer and colleagues (2018, p. 94) put it, “the ultimate aim of translational machine learning is to generate procedures that would be beneficial for clients, general practitioners, and in specialised hospital settings to improve patient outcomes”. This is a noble aim. However, the overall effect of viewing psychiatry this way is that discrete clinical decisions are foregrounded, while anything not amenable to being quantified and incorporated into a statistical learning problem is neglected. For example, the role of the psychiatrist as an empathic listener and the idiosyncratic content of mental distress, resistant to concise mathematical representation, can easily be obscured in this formulation (Horn & Weisz, 2020). Of course, these are relative effects. Function orientation does not make one blind to the interpersonal nature of psychiatry. However, to the extent that one develops expertise in a way of seeing that relies on noting and

measuring formalisable properties, one naturally becomes less attuned, on average, to those properties which cannot be easily formalised. This is what Cichy and Kaiser (2019, p. 312) mean when they stress that “models are not neutral tools.” In Chapter 6.4.1, I will argue that function orientation is a flawed perspective. Without addressing the limitations of currently available treatments, optimising clinical decision making cannot be transformative.

6.3 The Bottleneck of Understanding

Having characterised the data-driven approach, I now introduce the notion of *understanding* that I claim to be of critical importance to treating mental illness.¹⁰⁹ While this is a general account of understanding, when it is applied to how patients understand their own illnesses it takes on a crucial role for psychiatric science. In particular, I argue that, for patients suffering from mental illnesses, developing this form of understanding is necessary for recovery. According to my argument, patient understanding serves as a *bottleneck* which any effective approach to psychiatric science must address to improve patient outcomes.

6.3.1 Understanding as possession of adaptive mental constructs

I propose that understanding can be usefully conceived of as a set of adaptive mental constructs through which experience is mediated. I derive this definition by integrating ideas from theorists of psychiatry and the pragmatist tradition in philosophy. The notion of *understanding* has a rich history in psychiatry. Perhaps best known is Karl Jaspers’ account (1959/1997), on which understanding is a psychiatrist’s intuitive, empathic means of interpreting a patient’s mental world. However, while the interpersonal relationship between practitioner and patient is of immense importance to psychiatry, understanding can also be isolated and considered at an individual level. For example, consider German polymath Wilhelm Dilthey’s account (1894/1977). Central to Dilthey’s

¹⁰⁹ Note that I am not arguing that this is the *correct* or *best* way to conceptualise understanding, either across all contexts or even within the domain of psychiatry. There are other notions of understanding that may be valuable and important in different contexts. See Chapter 6.5.1 for more on this point.

notion of understanding is what he calls the “acquired nexus of psychic life” (1894/1977, p. 42). This nexus includes our values, habits, ideas, and goals and influences “every single act of consciousness” (1894/1977, p. 59). In Dilthey’s view, these mental constructs form a coherent whole that “determines the nature of our understanding of ourselves and of others” (1894/1977, p. 55). Following Dilthey, then, we will take the basis of understanding to consist of a set of mental constructs through which experience is mediated. However, we focus our attention specifically on patients’ understanding of their own inner lives, as this is what is of particular relevance to the goals of psychiatric science.

A further property that plays a role in conceptualisations of understanding from both inside and outside of psychiatry is that of *adaptivity*. In general, something is adaptive if it enables goals to be achieved under a set of environmental constraints. Within psychiatry, this notion appears in Aaron Beck’s cognitive model of mental illness (Beck, 1985; Beck & Haigh, 2014). On this view, a patient’s psychological distress is a downstream effect of their maladaptive *cognitive schemas*, which are the “beliefs, expectancies, evaluations, and attributions” that “serve to order everyday experience” (Beck & Haigh, 2014, p. 12). Note the similarity to Dilthey’s *acquired nexus*, described above. Under the cognitive model, interventions involve *cognitive restructuring*, which is the active modification of a patient’s maladaptive beliefs and schemas, to make them adaptive. However, one need not adopt the specific conceptual lexicon of the cognitive model in order to subscribe to the view that an adaptive notion of understanding is crucial to psychiatry. For example, Bruce Wampold defines adaptive explanations as those that “provide a means to overcome or cope with ... difficulties” (Wampold & Imel, 2015, p. 58). Wampold and Budge suggest that “patients typically present with a maladaptive explanation for their disorder” and “a primary therapeutic activity of the therapist is to provide an adaptive explanation” (2012, p. 612). Wampold’s work on the *common factors* of efficacious psychotherapy provides evidence that the delivery of such adaptive explanations is a necessary feature of *all* effective therapies (Wampold, 2015; Wampold & Imel, 2015).¹¹⁰ While this line of work focuses on adaptive *explanations*, from a patient’s

¹¹⁰ The idea that diverse methods of psychotherapy owe their efficacy to a set of shared common factors was first proposed by Saul Rosenzweig (1936) before being explored in greater depth by Jerome Frank (Frank

perspective, accepting and internalising such an explanation *just is* coming to understand their experience in an adaptive way.

The idea that understanding is usefully conceptualised as adaptive, affording particular abilities or skills, also resonates with recent currents of thought in the pragmatist philosophy of understanding (de Regt, 2019; Grimm, 2019; Hills, 2016; Wilkenfeld, 2013). Consider Daniel Wilkenfeld's account:

Understanding is at root the possession of the right sort of mental representations of that which is understood; [...] a mental representation counts as being "of the right sort" in virtue of the fact that possession of it enables one to perform [...] feats relevant in that context. (Wilkenfeld, 2013, p. 1000)

Notice that this view equates *understanding* with being empowered to *do* certain things. Namely, whatever is relevant to a particular context. Wilkenfeld's account is general but can be readily applied to psychiatric patients. The mental representations relevant to patients may involve aspects of self-understanding, life narratives, memories of and beliefs about aberrant experiences, and so on (Glover, 2020; McConnell, 2020). The relevant feats which the afflicted hope to be enabled to perform may be things like engagement in autonomous, competent behaviour, the pursuit of goals, and effective self-regulation, particularly in the face of stressors (Ivanov & Schwartz, 2021; Jacob, 2015).

So then, putting the ideas above together, patient understanding can be cast as the patient's possession of a set of adaptive mental constructs. Following Beck and Dilthey, I use the term *mental constructs* to refer to things like beliefs, habits, goals, expectancies, evaluations, values, and so on.¹¹¹ Following Wampold and Wilkenfeld, I use the term *adaptive* to mean enabling one to do relevant things in a given context. Notice two things about this notion of understanding. First, it covers both intuitive and intellectual senses of the word. By intuitive, I mean related to early, subconscious

& Frank, 1993). Various common factors, including the therapeutic alliance, empathy, and shared expectations, have been proposed (Wampold, 2015). While there is no consensus on a particular conceptual model, there is very strong evidence that the common factors account for most of the variance in clinical outcomes (Peterson, 2019).

¹¹¹ The argument does not turn on *which particular* mental constructs play the role of organizing ongoing experience. See Chapter 6.5.1 for further elaboration on this point.

aspects of information processing, like the biased emotional processing hypothesised to be important to depression (Harmer et al., 2009). By intellectual, I mean involving articulable propositional content, like a belief that one is incapable (Beck, 1985). Second, the notion of understanding we are using is not *factive*. That is, it does not require conformation to any particular standard of truth (Doyle et al., 2019). This is suitable for our context for two reasons. First, in the study of effective psychotherapy there is no empirical evidence that the truth value of the explanation is related to outcome (Frank & Frank, 1993; Wampold et al., 2007).¹¹² Second, for some of the mental constructs relevant to this notion of understanding, such as evaluations, it is not clear whether or how a truth value could be ascribed. I further clarify and defend this analysis of understanding in Chapter 6.5.1.

6.3.2 Understanding as a bottleneck to recovery

Above, we conceptualised understanding as possession of adaptive mental constructs. For those suffering from mental illness, gaining understanding might involve developing beliefs, expectancies, and values which enable the self-regulation of affective experience and behaviour. I now make the claim that *patient understanding is a bottleneck for robust recovery*.¹¹³

A bottleneck is a component within a broader system that limits the rate at which something is produced. For example, a shortage of semiconductor chips limits the rate at which cars can be manufactured. I suggest that patient understanding is a bottleneck for recovery in the following sense: the long-term rate at which a patient recovers is limited by the rate at which that patient develops understanding of their illness. In other words, the outcomes psychiatric science wants to bring about cannot be achieved without finding a way to reliably increase understanding. Including the ‘long-term’ qualifier captures the idea that short-term improvements in an outcome measure could reflect random fluctuations in environmental factors such as the presence and intensity

¹¹² Of course, such evidence may be forthcoming. There are theoretical reasons to expect some relationship between truth and efficacy, all else being equal (see Baker et al., 2008; Laska et al., 2014 for discussion of this issue).

¹¹³ I follow defenders of the *recovery model* in viewing recovery from mental illness as describing a *process towards a goal of living a self-efficacious and satisfying life* (Anthony, 2000; Jacob, 2015; Ramon et al., 2007).

of stressors. Such fluctuations are one of the reasons why recovery is typically a non-linear process that happens in “fits and starts” (Jacob, 2015, p. 118). Casting understanding as a bottleneck is a way of framing the idea that, given how we have defined understanding, it is *essential* for a robust recovery. If one lacks adaptive mental constructs, one cannot perform relevant feats in one’s own context. As previously stated, these feats may include setting and seeking personal goals, regulating affective experience, and engaging in a community. If one lacks the mental constructs that enable the performance of such feats, it is difficult to see how one could live a satisfying and self-efficacious life, conditions constitutive of recovery.¹¹⁴

I see the value of the bottleneck framing as being threefold. First, the concept of a bottleneck implies embeddedness in a system with multiple components. In this case, the system consists of psychiatric science as a whole, including disease classification, treatment development, therapeutic best practice, and so on. This makes the metaphor apt for sharpening thinking about the relationship between understanding and other levels of organisation which may be pertinent to improving outcomes. For example, variables specifiable at the genetic, behavioural, and social levels may each be critical intervention targets for specific instantiations of psychiatric disorder. That is, inducing the adaptive mental constructs that constitute understanding may, in some cases, be most effectively achieved through interventions that are primarily medicinal or behavioural, rather than psychological.¹¹⁵ However, the bottleneck framing highlights that in order to engender a robust recovery, *any* form of intervention must induce understanding. Beck (1985, pp. 333–334) makes a related point, noting that both psychotherapy and pharmacotherapy must alter defective cognitive content to be efficacious. Second, framing the challenge of psychiatry as addressing the bottleneck of understanding foregrounds subjective experience and psychopathology. Why is this a good thing? Because psychopathology is the very subject matter of psychiatry (Stanghellini & Broome, 2014). Neurobiological, genetic, and sociocultural approaches

¹¹⁴ While I have framed this in terms of the recovery model, the argument also applies, albeit less strongly, to symptom-focused models of recovery. To see this, consider that some of the feats relevant to patients will be constitutive of their diagnoses. For example, a patient with generalised anxiety disorder may hope to concentrate effectively on tasks and avoid becoming irritable (both symptoms of the disorder in the DSM5, American Psychiatric Association, 2013).

¹¹⁵ I thank an anonymous reviewer for pressing me to make this point clear.

are necessarily studying the correlates of what is, fundamentally, *mental* disorder (see Thornton, 2020, p. 235 for a more detailed version of this argument). Finally, focusing on a person-centered, intentional, mental ability like understanding when conceptualising the goals of psychiatry may be of pragmatic value. Both theory and empirical evidence suggest that a sense of agency and self-efficacy are important for recovery from mental illness (Glover, 2020; McConnell & Snoek, 2018). Framing the challenge of recovery as the development of understanding emphasises the power of the human subject to actively negotiate an alleviation of their own suffering (Anthony, 2000; Coulombe et al., 2016).

6.4 The Data-driven Approach and the Bottleneck of Understanding

If the view presented in Chapter 6.3 is correct, a transformative improvement to psychiatric science must involve unblocking the bottleneck of understanding. In this section I argue that the data-driven approach lacks a clear means of achieving this. First, I argue that, because extant treatments lack specificity, optimising their allocation will not address the bottleneck of understanding. Then, I argue that using the data-driven approach to target patient outcomes directly is fraught with theoretical challenges. Finally, I argue that there are conceptual difficulties plaguing the translation of data-driven insights into therapeutically valuable information that can aid patient understanding.

6.4.1 Limits of a function-oriented approach

Recall that the data-driven philosophy being *function-oriented* refers to how it conceptualises psychiatry as a collection of machine learning problems. As discussed in Chapter 6.2.3, methodological tools act like filters, rendering particular aspects of phenomena salient. In this case, the data-driven approach focuses on clinical decision-making under the implicit assumption that optimising these decisions will dramatically improve outcomes. Since clinical decisions can only be made with respect to existing interventions, the improvements the data-driven approach hopes to bring about must be realised through negotiating a more effective *allocation* of interventions (Bzdok & Meyer-

Lindenberg, 2018, p. 225). I call this the *indirect path* to transforming patient outcomes, since it does not involve targeting patient understanding directly.

The problem with the indirect path is the lack of *specificity* of extant treatments. A treatment would be specific if it had an active ingredient that rectified a particular dysfunction (Wampold & Imel, 2015, p. 60). Unfortunately, there is little evidence that treatments from either of the two major modalities, psychotherapy or pharmacotherapy, operate through specific ingredients that directly target pathology (Ivanov & Schwartz, 2021; Middleton & Moncrieff, 2019; Sathyanarayana Rao & Andrade, 2016; Wampold & Imel, 2015).¹¹⁶ In the case of psychotherapy, Wampold and Imel summarise their comprehensive review of the subject thusly, “there is no compelling evidence that the specific ingredients of any particular psychotherapy or specific ingredients in general are critical to producing the benefits of psychotherapy” (Wampold & Imel, 2015, p. 253). When it comes to pharmacotherapies, Middleton and Moncrieff state that “there is no evidence that antidepressants work by correcting a chemical imbalance or other identifiable abnormality” (Middleton & Moncrieff, 2019, p. 52). While there are many theories attempting to get a firmer handle on therapeutic mechanisms of action (Beck & Haigh, 2014; Harmer et al., 2017), these have not yet led to the development of specific treatments.

This lack of specificity in treatments is concerning for the data-driven approach. If it were the case that extant treatments were specific but misallocated, correcting their allocation would be a clear path to radically improved outcomes. However, given the lack of evidence that extant treatments are specific (Budd & Hughes, 2009; Wolpert et al., 2021), the effects of improved allocation may be small. The problems associated with mental disorders are, as Leichsenring and colleagues (Leichsenring et al., 2022, p. 141) put it when summarising their findings from a survey of meta-analyses across major mental disorders, “not sufficiently addressed by the available treatments”. By analogy, imagine a restaurant full of diners randomly assigned variously shaped spoons to cut thick steaks. An optimal reallocation of the spoons according to hand size would be unlikely to bring about a transformative improvement in the average carving ability of the

¹¹⁶ As Wampold and Imel state (Wampold & Imel, 2015, p. 47), the necessity of some form of exposure for treating specific phobias may be an exception to this rule.

patrons. The aim of data-driven psychiatry can be summarised as the prescription of the right treatment to the right person at the right time (Trivedi, 2016). However, if a specific treatment has not been devised yet, no algorithm can recommend its prescription.

6.4.2 Limits of a theory-neutral approach

Above we questioned the *indirect path* by which the data-driven approach could revolutionise patient outcomes. What about a *direct path*? The idea here would be that the data-driven approach could improve treatments by optimising for patient understanding itself. In other words, machine learning could be used to tailor treatments to achieve better outcomes. Research in this area is already underway (see Chekroud et al., 2021; Su et al., 2020 for reviews). In one study, researchers coded therapist utterances in cognitive behavioural therapy sessions and trained a model to determine the association between therapeutic content and clinical outcomes (Ewbank et al., 2020). In this case, the quantity of utterances classed as cognitive and behavioural change methods in therapy sessions was found to correlate with better chances of symptom improvement. While such early results are valuable, notice the theoretical challenges involved in pushing such research programs further. Ewbank and colleagues used a cognitive behavioural therapy framework to define twenty-four classes of utterance (e.g., *mood check*, *set goals*, *therapeutic empathy*). The coding of these utterances runs into thorny theory-laden issues of variability and interpretation (Horn & Weisz, 2020). What are the right categories? How many categories can a single utterance belong to? Should semantically distinct ways of expressing similar sentiments be considered therapeutically equivalent? Far from the theory-neutral philosophy of the data-driven approach, applying the methodology to improving treatment demands introducing significant theoretical assumptions. This should not be read as a criticism of this research direction. Rather, it is an attempt to show that for the data-driven approach to tackle understanding, interdisciplinarity and theory are required. The hypothesis-free number crunching espoused by some proponents of the data-driven approach is not a viable option.

There is a different kind of problem for the strategy of developing pharmacological treatments to overcome the lack of specificity mentioned in Chapter 6.4.1. In Chekroud and colleagues' review of data-driven approaches to predicting treatment outcomes they

find that factors at the psychological and social levels “consistently offer more meaningful and generalisable predictions” (2021, p. 165) than those at the neurobiological and genetic levels. In other words, when it comes to treatment outcomes, most of the predictive signal is not to be found at a level that can be targeted directly by pharmacological interventions. This means that, without a well-validated mapping between the predictive psychological constructs and the neurobiological entities that implement them, pharmacological interventions cannot be targeted directly at understanding (Ivanov & Schwartz, 2021; Poldrack & Yarkoni, 2016). As Ivanov and Schwartz (2021, p. 5) say, “we are still very far indeed” from possessing such a mapping. This critique is not meant to trivialise the important work of using the data-driven approach to improve the allocation of extant treatments for conditions like depression (Chekroud et al., 2016; Nie et al., 2018). However, these success stories are of the modest, indirect variety. They do not give any reason to think that the data-driven approach can optimise novel pharmacological treatments to directly target psychopathology.

6.4.3 Translation problems for data-driven insights

Another avenue by which the data-driven approach might transform mental healthcare is by extracting therapeutically actionable insights from trained models. Indeed, a common assertion among proponents of the data-driven approach is that these methods might discover subtle and complex patterns which aid mechanistic understanding of mental disorders. For example, Durstewitz and colleagues (2019, p. 1592) suggest that as well as using deep neural networks for prediction, we may be able “to employ them to also gain insight into physiological, computational, and cognitive mechanisms”. Along the same lines, Koppe and colleagues (2021, p. 187) state that using visualisation techniques on machine learning models “may help uncover interpretable multi-modal biomarkers of psychiatric disease”.

One reason to think this might be so is the *convergence logic* that characterises data-driven success stories in science. I introduce the term convergence logic to refer to the idea that, because accurate prediction in complex, high-dimensional domains is exceedingly difficult, any model that attains high accuracy under tight constraints *must* do so in virtue of learning robust patterns which capture core features of the data domain

being modelled (Schrimpf et al., 2020). This line of reasoning is behind the neuroscientific research program using deep neural networks to model the primate ventral visual stream (Yamins & DiCarlo, 2016). In that case, the conjecture is that in order to be able to accurately classify images, a neural network *has to* converge upon mechanisms that approximate its biological target.¹¹⁷ The convergence logic is also manifest in the linguistic structure that is learned by neural network language models (Manning et al., 2020). In order for a model to accurately predict language tokens, it *has to* internalise aspects of linguistic structure. Moving back to psychiatry, if large computational models incorporating multiple modalities are able to generate accurate prognoses or reliably pick out novel disorder subtypes, one might think *it must be* because they are converging on robust structure in the data.¹¹⁸ On this view, machine learning models might be the right tools to discover the kind of “fuzzy, cross-level mechanisms” that Kendler suggests might be the “true nature of psychiatric disorders” (2012, p. 18).

How would such discovery relate to patient understanding? Kendler and Campbell (2014, p. 1) suggest that scientific knowledge might “expand the domain of the understandable” by providing models of psychopathological processes that render them intelligible (note that Kendler and Campbell are referring to Jaspers’ clinician-centered notion of understanding here). As an example, the authors discuss the pathological misattribution of meaning that leads to delusional beliefs in schizophrenia, considered paradigmatically un-understandable by Jaspers. Modern neuroscience posits the aberrant behaviour of dopaminergic neurons encoding motivational salience to explain how these beliefs are formed. Kendler and Campbell suggest that this form of dopamine system dysfunction *makes sense of* the psychopathological experiences. If the data-driven approach could provide scientific knowledge about a range of psychiatric disorders, those insights might be used to provide adaptive explanations to patients and induce the understanding required to radically improve outcomes.¹¹⁹ Of course, there

¹¹⁷ Naturally, this holds only for a given level of abstraction.

¹¹⁸ We are assuming here that rigorous cross-validation rules out models achieving accuracy by overfitting on spurious correlations.

¹¹⁹ Other authors have argued that bringing neuroscience into the clinic could have negative consequences, inducing fatalism by conveying a deterministic world, stripped of human agency (Hyman & McConnell, 2020).

may be fundamental limits to this enterprise in certain cases. For example, it is difficult to see how scientific understanding could translate into an understanding of psychopathological experiences which lack internal coherence.¹²⁰ This is an avenue that future research must explore. However, even if the convergence logic holds and novel multilevel patterns which *can* be mapped onto features of mental disorder are discovered, there are still important barriers for translating data-driven insights into improved clinical outcomes.¹²¹

The first is the issue of intelligibility. The flexibility and expressivity of the machine learning paradigm come with a cost. *How* these models transform input into output is, *prima facie*, unintelligible (Zednik, 2021). In their review of machine learning applications in psychiatry, Cearns and colleagues (2019, p. 9) claim forcefully that “knowing every single one of the hundreds of millions of parameter values in a given ML model would fail to provide even a spec of practically useful insight into the inner workings of a trained model”. However, this assessment is overly pessimistic. While it is clear that no human could infer meaning directly from hundreds of millions of parameter values, functional characterisation of trained models at a higher-level of abstraction, such as layers or pathways in a neural network, may be possible. In the domain of psychiatry, this would require translating the components of trained machine learning models into psychiatrically meaningful terms (Sheu, 2020). Strategies for interpreting the internal structure of trained models are being developed (Elhage et al., 2021; Koppe et al., 2021; Olah et al., 2018; Sheu, 2020). Researchers following such strategies, such as Anthropic’s *mechanistic interpretability* approach (Elhage et al., 2021), use a combination of mathematics and experimental techniques to probe machine learning models in an attempt to characterise their functional organisation. While further research is clearly required, applying interpretability techniques to large predictive models is a promising option for discovering the multi-level structure of psychiatric disorders.

¹²⁰ I thank an anonymous reviewer for raising this point.

¹²¹ There is not space here to comment on whether such patterns would necessarily constitute mechanisms. The argument does not turn on whether or not they would.

Unfortunately, even the discovery of novel insights into mental illness would not guarantee a transformative impact for the data-driven approach. This is due to what I call the *translation problem*. Namely, that the understanding gleaned by psychiatrists and the scientific community still needs to be translated into therapeutic techniques that induce *patient* understanding. If a patient cannot internalise an explanation, then that explanation will not be adaptive for them. Explanations invoking fuzzy and complex multi-level mechanisms, even if understood by psychiatrists, may not be communicable to the average patient. In order for these hard-won insights to be clinically relevant, they need to be translated into language that resonates with the “assumptive worlds” of specific patients (Frank & Frank, 1993, p. 30).¹²² Supporting this point, research has shown that having a *shared rationale* with a practitioner is important to patients and contributes to improved outcomes (Cuevas et al., 2014; Johansson & Eklund, 2003; Wampold & Budge, 2012). Further, non-specific components of psychiatric treatments, such as clinical communication and treatment framing, also play a significant role in patient outcomes (Priebe et al., 2020). This suggests that how explanations are presented is important and that the inferential distance between practitioner and patient should not be stretched too far. These points highlight the difficulty of the translation problem. Discovering the multi-level structure of psychiatric disorder is not enough. The data-driven approach still requires important theoretical work to be done before the bottleneck of understanding can be addressed.

6.4.4 Implications

Overall, then, while the data-driven approach is a fascinating and powerful addition to the landscape of psychiatric science, it has no straightforward means of unblocking the bottleneck of understanding. As such, the suggestion that this approach will bring about a transformative improvement to patient outcomes is not warranted. This is not to claim that the data-driven approach has no role to play. Though I have argued it will not be transformative, increasing the efficiency of the allocation of extant treatments is still an

¹²² Wampold and colleagues flesh this notion out by suggesting that explanations “should be proximal to the client’s currently held explanation or expectation, and should not create dissonance with the attitudes and values of the client” (Wampold et al., 2007, p. 125). This coheres with the point, explained in Chapter 6.3.1, that adaptive explanations, and by extension understanding, need not be true.

important goal. Further, the data-driven analysis of pharmacological and psychotherapeutic interventions is a promising line of research from which important principles and insights regarding treatment strategies can be derived (Ewbank et al., 2021). Finally, discoveries about how fine-grained patterns of variation across multiple levels relate to psychopathology and the course of mental illness will surely contribute to how we understand and treat psychiatric disorder. To be clear then, the conclusion of the argument is that the contribution of the data-driven approach ought to be incremental and interdisciplinary, rather than discontinuous and technocratic. Proponents take pains to acknowledge the technical, ethical, and practical challenges facing their endeavour (Bzdok & Meyer-Lindenberg, 2018; Cearnns et al., 2019). However, the function-oriented and theory-neutral philosophy often obscures the deep *conceptual* challenges pervading a data-driven approach to psychiatric science. The problem of inducing understanding in patients suffering mental distress cannot be solved by data alone.

6.5 Responses to Objections

I now anticipate and respond to possible objections to two of the core concepts invoked in my argument.

6.5.1 On understanding

I cast *understanding* as the possession of adaptive mental constructs that enable a subject to perform feats relevant to them. However, critics might protest that this pragmatic notion has not been adequately defended. I will briefly address three potential worries. First, there are many competing conceptualisations of understanding out there, so why pick this one in particular? Second, does this notion really capture the views of the scholars I cited? Third, does this conceptualisation really capture anything unified at all?

On the first issue, as to why I choose this notion of understanding rather than another, I make a clarification rather than supplying an argument. I endorse a form of conceptual pluralism whereby concepts can be invoked or defined stipulatively for the sake of argument so long as this practice is not abused to wilfully obfuscate. The concept I am invoking was inspired by notions of understanding from pragmatist philosophy and

theoretical work in psychiatry. I take this to suffice to show that my usage is neither without precedent nor needlessly idiosyncratic. Ultimately, for the purposes of this paper, whether *understanding* is the right term to label the concept I have described is less important than whether or not that concept plays the role I have suggested in limiting the potential of the data-driven approach to psychiatry.

A more serious worry is that I have erroneously conflated distinct views and presented them under a common label. Along these lines, one might worry that the *acquired nexus* invoked by Wilhelm Dilthey (1894/1977) in his discussion of understanding and the *cognitive schemas* central to Aaron Beck's (Beck & Haigh, 2014) theorising are rather distinct from the *adaptive explanations* that Wampold (Wampold & Budge, 2012) describes. The former two notions involve subconscious aspects of mentality that organise ongoing perception and experience, whereas the latter notion seems to be talking about explicit and propositional explanations. Are these authors not pointing to distinct concepts? I think a close reading shows that they each have something broader in mind. First, both Dilthey and Beck include propositional entities like beliefs when spelling out their ideas. They do not limit their concepts to the unconscious. Second, Wampold takes a very liberal stance on what constitutes an explanation. For example, he describes an inability to recognise one has a problem as a dysfunctional explanation, even though this would clearly not be an articulable proposition (Wampold et al., 2007, p. 123). As such, despite their different framings, these authors are capturing a broadly co-extensive set of mental constructs. Namely, those that play a critical role in structuring, organising, and influencing ongoing experience.

Of course, even if I am right that *understanding* captures what Dilthey, Beck, and Wampold have in mind, it may still be the case that it does not play the role I am claiming it does. A third objection that one might have along these lines is that this notion of understanding is both too broad and too vague to be useful. It is broad in the sense described above, covering both subconscious and conscious constructs. It is vague in the sense that I have not specified precisely what set of things is supposed to be captured by *mental constructs*. In response to this worry, I stress that both the broadness of the notion and the vagueness with respect to what kind of entities are being referred to are features of the account, not bugs. Much of my argument turns on the fact that we have

an inadequate theory of *how exactly* these influences – beliefs, values, goals, expectations, and the like – give rise to psychopathological experience. If we knew which constructs were relevant, how they interact with one another, and how to specify and measure them, the prospects of the data-driven approach addressing the bottleneck of understanding would be significantly improved (see Poldrack & Yarkoni, 2016 for an exposition of the view that there is unlikely to be any single ‘correct’ way to carve psychology into a set of mental constructs). As such, I gladly acknowledge that my account is underspecified. In my view, this reflects an appropriate degree of epistemic uncertainty with respect to the details.

6.5.2 On the bottleneck metaphor

One possible objection to casting understanding as a bottleneck is that the mental states that make up understanding do not bear the appropriate relation to the other components in the system to justify the metaphor. In particular, a proponent of the biological approach to psychiatry might argue that a suitable pharmacological intervention not only alters a patient’s brain states, but also alters their mental states. This interlocuter could suggest that treating the underlying neurobiology and the mental dysfunction as independent smacks of Cartesian dualism. On this view, regardless of the messy details, as long as subjective psychopathology is *implemented* by neural wetware, it stands to reason that intervening on that wetware *constitutes* intervening on the patient’s relevant mental constructs. This threatens the applicability of the bottleneck metaphor, which implies that pharmacological interventions are causally upstream of the mental constructs that feature in understanding.

It is certainly true that the complex relation holding between brain and mind makes psychiatry significantly trickier to reason about than other systems which experience bottlenecks, such as factory production lines. However, for the bottleneck metaphor to be useful, all that is required is a dissociation between the neural changes known to be induced by pharmacological agents and the development of understanding. Notice that such a dissociation is a function of our state of knowledge, and thus not a strong claim about the relation between brain and mind. The utility of treating the cognitive and chemical as dissociable can be seen in the cognitive neuropsychological theory of how antidepressants work (Harmer et al., 2009, 2017). On this view, the early biological

effects of selective serotonin reuptake inhibitors lead to a positive shift in the processing of emotionally salient information, which then leads to improved mood. This case highlights how scholars still find it useful to consider the chemical and the cognitive as functionally distinct. Of course, Harmer and colleagues are well aware that the cognitive effects they point to are implemented by further biological changes (such as increased neural plasticity). The point is that the precise details of these changes are not yet known and hence cannot be targeted directly by pharmacological intervention (see also Chapter 6.4.2). Further, it has been shown that no specific feature of brain biology is sufficient to demarcate healthy and pathological populations (First et al., 2018; Holmes & Patrick, 2018). This implies that the dissociation should hold for any pharmacological agent which targets a particular biomarker, at least to the extent that there is uncertainty regarding the agent's downstream effects. Of course, it is possible that future technology will enable fundamentally different kinds of neurobiological intervention. However, this does not threaten the bottleneck framing, which is specifically intended to characterise the challenge psychiatric science faces *today*, given our current epistemic and technological situation. In sum, I maintain it is reasonable to consider patient understanding a bottleneck to robust recovery.

6.6 Conclusion

The focus of this article is the question: Does the data-driven approach have transformative potential for psychiatric science? In providing an answer to this question, I focused on a core limitation. Namely, that *patient understanding* functions as a bottleneck for recovery, and the data-driven approach is ill-suited to address that bottleneck. If this view is correct, talk of an AI revolution in mental healthcare is premature. Given the current state of psychiatric science, the efficient allocation of treatments, which the data-driven approach can deliver, is likely to yield only modest improvements to mental health outcomes. In order to unlock additional benefits, further knowledge of the specificities of psychopathology and how treatments can be tailored to induce the understanding necessary for recovery are needed. This may involve engaging with research on features of psychiatry that are difficult to formalise, such as empathy, non-verbal communication, and the framing of treatments (Priebe et al., 2020), the role

of placebo effects in treatment outcomes (Enck & Zipfel, 2019), and the practitioner-patient relationship (Horn & Weisz, 2020). As explained in Chapter 6.4.2, this line of research involves many researcher degrees of freedom and cannot proceed in a theory-neutral way. While the data-driven approach can contribute to these endeavours, it will do so best in collaboration with a theory-driven attempt to understand mental disorders.

6.7 References

- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders*. DSM Library.
<https://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596>
- Anthony, W. A. (2000). A recovery-oriented service system: Setting some system level standards. *Psychiatric Rehabilitation Journal*, 24(2), 159–168.
<https://doi.org/10.1037/h0095104>
- Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145(Pt B), 137–165.
<https://doi.org/10.1016/j.neuroimage.2016.02.079>
- Baker, T. B., McFall, R. M., & Shoham, V. (2008). Current Status and Future Prospects of Clinical Psychology: Toward a Scientifically Principled Approach to Mental and Behavioral Health Care. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 9(2), 67–103. <https://doi.org/10.1111/j.1539-6053.2009.01036.x>
- Balaskas, A., Schueller, S. M., Cox, A. L., & Doherty, G. (2021). Ecological momentary interventions for mental health: A scoping review. *PLOS ONE*, 16(3), e0248152.
<https://doi.org/10.1371/journal.pone.0248152>
- Battaglia, P., Hamrick, J. B. C., Bapst, V., Sanchez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K., Nash, C., Langston, V. J., ... Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv*.
<https://arxiv.org/pdf/1806.01261.pdf>
- Beck, A. T. (1985). Cognitive Therapy, Behavior Therapy, Psychoanalysis, and Pharmacotherapy. In M. J. Mahoney & A. Freeman (Eds.), *Cognition and Psychotherapy* (pp. 325–347). Springer US. https://doi.org/10.1007/978-1-4684-7562-3_14
- Beck, A. T., & Haigh, E. A. P. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual Review of Clinical Psychology*, 10, 1–24.
<https://doi.org/10.1146/annurev-clinpsy-032813-153734>
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. <https://doi.org/10.1109/tpami.2013.50>

Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The Two Cultures of Computational Psychiatry. *JAMA Psychiatry*, 76(6), 563–564.

<https://doi.org/10.1001/jamapsychiatry.2019.0231>

Bickman, L. (2020). Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. *Administration and Policy in Mental Health*, 47(5), 795–843.

<https://doi.org/10.1007/s10488-020-01065-8>

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 42, e2. <https://doi.org/10.1017/S0140525X17002266>

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [Cs, Stat]*.

<http://arxiv.org/abs/2104.13478>

Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry*, 218(3), 131–134. <https://doi.org/10.1192/bjp.2019.245>

Budd, R., & Hughes, I. (2009). The Dodo Bird Verdict--controversial, inevitable and important: A commentary on 30 years of meta-analyses. *Clinical Psychology & Psychotherapy*, 16(6), 510–522. <https://doi.org/10.1002/cpp.648>

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3. <https://doi.org/10.1016/j.bpsc.2017.11.007>

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, 9(1), Article 1.

<https://doi.org/10.1038/s41398-019-0607-2>

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 20(2), 154–170.

<https://doi.org/10.1002/wps.20882>

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet. Psychiatry*, 3(3), 243–250. [https://doi.org/10.1016/S2215-0366\(15\)00471-X](https://doi.org/10.1016/S2215-0366(15)00471-X)

Chollet, F. (2021). *Deep Learning with Python, Second Edition*. Simon and Schuster.

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4), 305–317. <https://doi.org/10.1016/j.tics.2019.01.009>

Coulombe, S., Radziszewski, S., Meunier, S., Provencher, H., Hudon, C., Roberge, P., Provencher, M. D., & Houle, J. (2016). Profiles of Recovery from Mood and Anxiety Disorders: A Person-Centered Exploration of People's Engagement in Self-Management. *Frontiers in Psychology*, 7.

<https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00584>

Cuevas, C. D. las, Peñate, W., & Rivera, L. de. (2014). To what extent is treatment adherence of psychiatric patients influenced by their participation in shared decision

making? *Patient Preference and Adherence*, 8, 1547–1553.

<https://doi.org/10.2147/PPA.S73029>

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, 11(1), 126. <https://doi.org/10.1186/1741-7015-11-126>

de Boer, N. S., de Bruin, L. C., Geurts, J. J. G., & Glas, G. (2021). The Network Theory of Psychiatric Disorders: A Critical Assessment of the Inclusion of Environmental Factors. *Frontiers in Psychology*, 12, 221. <https://doi.org/10.3389/fpsyg.2021.623970>

de Regt, H. W. (2019). From Explanation to Understanding: Normativity Lost? *Journal for General Philosophy of Science*, 50(3), 327–343. <https://doi.org/10.1007/s10838-019-09477-3>

Dembić, S. (2023). Mental disorder: An ability-based view. *Philosophy and the Mind Sciences*, 4. <https://doi.org/10.33735/phimisci.2023.9630>

Department of Health and Social Care. (2021, March 27). *Mental health recovery plan backed by £500 million*. GOV.UK. <https://www.gov.uk/government/news/mental-health-recovery-plan-backed-by-500-million>

Dickinson, E. (1998). *The Poems of Emily Dickinson*. Harvard University Press. (Original work published 1890)

Dilthey, W. (1977). Ideas Concerning a Descriptive and Analytic Psychology (1894) (R. M. Zaner, Trans.). In W. Dilthey (Ed.), *Descriptive Psychology and Historical Understanding* (pp. 21–120). Springer Netherlands. https://doi.org/10.1007/978-94-009-9658-8_2 (Original work published 1894)

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>

Doyle, Y., Egan, S., Graham, N., & Khalifa, K. (2019). Non-factive Understanding: A Statement and Defense. *Journal for General Philosophy of Science*, 50(3), 345–365. <https://doi.org/10.1007/s10838-019-09469-3>

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, 24(11), Article 11. <https://doi.org/10.1038/s41380-019-0365-9>

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, 14, 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., ... Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.

Elujide, I., Fashoto, S. G., Fashoto, B., Mbunge, E., Folorunso, S. O., & Olamijuwon, J. O. (2021). Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. *Informatics in Medicine Unlocked*, 23, 100545. <https://doi.org/10.1016/j.imu.2021.100545>

Enck, P., & Zipfel, S. (2019). Placebo Effects in Psychotherapy: A Framework. *Frontiers in Psychiatry*, 10. <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00456>

- Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science (New York, N.Y.)*, 196(4286), 129–136. <https://doi.org/10.1126/science.847460>
- Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*, 77(1), 35–43. <https://doi.org/10.1001/jamapsychiatry.2019.2664>
- Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3), 300–312. <https://doi.org/10.1080/10503307.2020.1788740>
- Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, 23(7), 584–601. <https://doi.org/10.1016/j.tics.2019.03.009>
- Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of ‘precision psychiatry’. *BMC Medicine*, 15(1), 80. <https://doi.org/10.1186/s12916-017-0849-x>
- First, M. B., Rebello, T. J., Keeley, J. W., Bhargava, R., Dai, Y., Kulygina, M., Matsumoto, C., Robles, R., Stona, A.-C., & Reed, G. M. (2018). Do mental health professionals use diagnostic classifications the way we think they do? A global survey. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 17(2), 187–195. <https://doi.org/10.1002/wps.20525>
- Frances, A. (2009). Whither DSM-V? *The British Journal of Psychiatry: The Journal of Mental Science*, 195(5), 391–392. <https://doi.org/10.1192/bjp.bp.109.073932>
- Frank, J. D., & Frank, J. B. (1993). *Persuasion and Healing: A Comparative Study of Psychotherapy*. JHU Press.
- García-Gutiérrez, M. S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A., & Manzanares, J. (2020). Biomarkers in Psychiatry: Concept, Definition, Types and Relevance to the Clinical Reality. *Frontiers in Psychiatry*, 11, 432. <https://doi.org/10.3389/fpsy.2020.00432>
- Gauld, C., Micoulaud-Franchi, J.-A., & Dumas, G. (2021). Comment on Starke et al.: ‘Computing schizophrenia: ethical challenges for machine learning in psychiatry’: from machine learning to student learning: pedagogical challenges for psychiatry. *Psychological Medicine*, 51(14), 2509–2511. <https://doi.org/10.1017/S0033291720003906>
- Ghaemi, S. N. (2007). Existence and pluralism: The rediscovery of Karl Jaspers. *Psychopathology*, 40(2), 75–82. <https://doi.org/10.1159/000098487>
- Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, 18, 34–42. <https://doi.org/10.1016/j.cobeha.2017.07.003>
- Glover, J. (2020). Psychiatry, folk psychology, and the impact of neuroscience—A response to Steve Hyman’s Loebel Lectures. In editor J. Savulescu, R. Roache, W. Davies, & editor J. P. Loebel (Eds.), *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 301–319). Oxford University Press.

- Gould, I. C., Shepherd, A. M., Laurens, K. R., Cairns, M. J., Carr, V. J., & Green, M. J. (2014). Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach. *NeuroImage: Clinical*, 6, 229–236. <https://doi.org/10.1016/j.nicl.2014.09.009>
- Grimm, S. R. (Ed.). (2019). *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*. Oxford University Press.
<https://doi.org/10.1093/oso/9780190860974.001.0001>
- Harmer, C. J., Duman, R. S., & Cowen, P. J. (2017). How do antidepressants work? New perspectives for refining future treatment approaches. *The Lancet. Psychiatry*, 4(5), 409–418. [https://doi.org/10.1016/S2215-0366\(17\)30015-9](https://doi.org/10.1016/S2215-0366(17)30015-9)
- Harmer, C. J., Goodwin, G. M., & Cowen, P. J. (2009). Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *The British Journal of Psychiatry*, 195(2), 102–108.
<https://doi.org/10.1192/bjp.bp.108.051193>
- Hengartner, M. P., & Lehmann, S. N. (2017). Why Psychiatric Research Must Abandon Traditional Diagnostic Classification and Adopt a Fully Dimensional Scope: Two Solutions to a Persistent Problem. *Frontiers in Psychiatry*, 8, 101.
<https://doi.org/10.3389/fpsy.2017.00101>
- Hills, A. (2016). Understanding Why. *Noûs*, 50(4), 661–688.
<https://doi.org/10.1111/nous.12092>
- Holmes, A. J., & Patrick, L. M. (2018). The Myth of Optimality in Clinical Neuroscience. *Trends in Cognitive Sciences*, 22(3), 241–257. <https://doi.org/10.1016/j.tics.2017.12.006>
- Horn, R. L., & Weisz, J. R. (2020). Can Artificial Intelligence Improve Psychotherapy Research and Practice? *Administration and Policy in Mental Health and Mental Health Services Research*, 47(5), 852–855. <https://doi.org/10.1007/s10488-020-01056-9>
- Hyman, S. E., & McConnell, D. (2020). Mental illness: The collision of meaning with mechanism. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 263–289). Oxford University Press.
<https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697>
- Insel, T. R., Collins, P. Y., & Hyman, S. E. (2015). Darkness Invisible: The Hidden Global Costs of Mental Illness. *Foreign Affairs*, 94(1), 127–135.
- Ivanov, I., & Schwartz, J. M. (2021). Why Psychotropic Drugs Don't Cure Mental Illness—But Should They? *Frontiers in Psychiatry*, 12.
<https://www.frontiersin.org/article/10.3389/fpsy.2021.579566>
- Jacob, K. S. (2015). Recovery Model of Mental Illness: A Complementary Approach to Psychiatric Care. *Indian Journal of Psychological Medicine*, 37(2), 117–119.
<https://doi.org/10.4103/0253-7176.155605>
- Jaspers, K. (1997). *General Psychopathology* (J. Hoenig & M. Hamilton, Trans.). JHU Press. (Original work published 1959)
- Johansson, H., & Eklund, M. (2003). Patients' opinion on what constitutes good psychiatric care. *Scandinavian Journal of Caring Sciences*, 17(4), 339–346.
<https://doi.org/10.1046/j.0283-9318.2003.00233.x>

- Joober, R., & Tabbane, K. (2019). From the neo-Kraepelinian framework to the new mechanical philosophy of psychiatry: Regaining common sense. *Journal of Psychiatry & Neuroscience : JPN*, 44(1), 3–7. <https://doi.org/10.1503/jpn.180240>
- Kendler, K. S. (2012). Levels of explanation in psychiatric and substance use disorders: Implications for the development of an etiologically based nosology. *Molecular Psychiatry*, 17(1), 11–21. <https://doi.org/10.1038/mp.2011.70>
- Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, 15(1), 5–12. <https://doi.org/10.1002/wps.20292>
- Kendler, K. S., & Campbell, J. (2014). Expanding the domain of the understandable in psychiatric illness: An updating of the Jasperian framework of explanation and understanding. *Psychological Medicine*, 44(1), 1–7. <https://doi.org/10.1017/S0033291712003030>
- Kendler, K. S., & Gyngell, C. (2020). Multilevel interactions and the dappled causal world of psychiatric disorders. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 25–45). Oxford University Press. <https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697>
- Keogh, E., & Mueen, A. (2017). Curse of Dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_192
- Koppe, G., Meyer-Lindenberg, A., & Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, 46(1), Article 1. <https://doi.org/10.1038/s41386-020-0767-z>
- Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., ... Meisenzahl, E. (2021). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry*, 78(2), 1–15. <https://doi.org/10.1001/jamapsychiatry.2020.3604>
- Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1), 33–67. <https://doi.org/10.1007/s13194-012-0056-8>
- Laska, K. M., Gurman, A. S., & Wampold, B. E. (2014). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy (Chicago, Ill.)*, 51(4), 467–481. <https://doi.org/10.1037/a0034332>
- Leder, G., & Zawidzki, T. (2023). The skill of mental health: Towards a new theory of mental health and disorder. *Philosophy and the Mind Sciences*, 4. <https://doi.org/10.33735/phimisci.2023.9684>
- Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: An umbrella review and meta-analytic evaluation of recent meta-analyses. *World Psychiatry*, 21(1), 133–145. <https://doi.org/10.1002/wps.20941>

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054. <https://doi.org/10.1073/pnas.1907367117>

McConnell, D. (2020). The proper place of subjectivity, meaning, and folk psychology in psychiatry. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 290–303). Oxford University Press. <https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697>

McConnell, D., & Snoek, A. (2018). The Importance of Self-Narration in Recovery From Addiction. *Philosophy, Psychiatry, and Psychology*, 25(3), 31–44. <https://doi.org/10.1353/ppp.2018.0022>

McCoy, L. G., Nagaraj, S., Morgado, F., Harish, V., Das, S., & Celi, L. A. (2020). What do medical students actually need to know about artificial intelligence? *Npj Digital Medicine*, 3(1), Article 1. <https://doi.org/10.1038/s41746-020-0294-7>

McGorry, P. D., & Nelson, B. (2019). Transdiagnostic psychiatry: Premature closure on a crucial pathway to clinical utility for psychiatric diagnosis. *World Psychiatry*, 18(3), 359–360. <https://doi.org/10.1002/wps.20679>

Middleton, H., & Moncrieff, J. (2019). Critical psychiatry: A brief overview. *BJPsych Advances*, 25(1), 47–54. <https://doi.org/10.1192/bja.2018.38>

Miranda, L., Paul, R., Pütz, B., Koutsouleris, N., & Müller-Myhsok, B. (2021). Systematic Review of Functional MRI Applications for Psychiatric Disease Subtyping. *Frontiers in Psychiatry*, 12, 665536. <https://doi.org/10.3389/fpsy.2021.665536>

Najafpour, Z., Fatemi, A., Goudarzi, Z., Goudarzi, R., Shayanfard, K., & Noorizadeh, F. (2021). Cost-effectiveness of neuroimaging technologies in management of psychiatric and insomnia disorders: A meta-analysis and prospective cost analysis. *Journal of Neuroradiology = Journal De Neuroradiologie*, 48(5), 348–358. <https://doi.org/10.1016/j.neurad.2020.12.003>

Nie, Z., Vairavan, S., Narayan, V. A., Ye, J., & Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PLOS ONE*, 13(6), e0197268. <https://doi.org/10.1371/journal.pone.0197268>

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010. <https://doi.org/10.23915/distill.00010>

Paulus, M. P. (2015). Pragmatism Instead of Mechanism: A Call for Impactful Biological Psychiatry. *JAMA Psychiatry*, 72(7), 631–632. <https://doi.org/10.1001/jamapsychiatry.2015.0497>

Pelin, H., Ising, M., Stein, F., Meinert, S., Meller, T., Brosch, K., Winter, N. R., Krug, A., Leenings, R., Lemke, H., Nenadić, I., Heilmann-Heimbach, S., Forstner, A. J., Nöthen, M. M., Opel, N., Reppe, J., Pfarr, J., Ringwald, K., Schmitt, S., ... Andlauer, T. F. M. (2021). Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 46(11), 1895–1905. <https://doi.org/10.1038/s41386-021-01051-0>

- Peterson, B. S. (2019). Editorial: Common factors in the art of healing. *Journal of Child Psychology and Psychiatry*, 60(9), 927–929. <https://doi.org/10.1111/jcpp.13108>
- Poldrack, R. A., & Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annual Review of Psychology*, 67(1), 587–612. <https://doi.org/10.1146/annurev-psych-122414-033729>
- Priebe, S., Conneely, M., McCabe, R., & Bird, V. (2020). What can clinicians do to improve outcomes across psychiatric treatments: A conceptual review of non-specific components. *Epidemiology and Psychiatric Sciences*, 29. <https://doi.org/10.1017/S2045796019000428>
- Qureshi, M. N. I., Min, B., Jo, H. J., & Lee, B. (2016). Multiclass Classification for the Differential Diagnosis on the ADHD Subtypes Using Recursive Feature Elimination and Hierarchical Extreme Learning Machine: Structural MRI Study. *PLOS ONE*, 11(8), e0160697. <https://doi.org/10.1371/journal.pone.0160697>
- Ramon, S., Healy, B., & Renouf, N. (2007). Recovery from Mental Illness as an Emergent Concept and Practice in Australia and the UK. *International Journal of Social Psychiatry*, 53(2), 108–122. <https://doi.org/10.1177/0020764006075018>
- Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, 6, 412–415. <https://doi.org/10.1111/j.1939-0025.1936.tb05248.x>
- Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: Toward clinical applications. *Current Opinion in Neurobiology*, 55, 152–159. <https://doi.org/10.1016/j.conb.2019.02.006>
- Sathyanarayana Rao, T. S., & Andrade, C. (2016). Classification of psychotropic drugs: Problems, solutions, and more problems. *Indian Journal of Psychiatry*, 58(2), 111–113. <https://doi.org/10.4103/0019-5545.183771>
- Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3), 413–423. <https://doi.org/10.1016/j.neuron.2020.07.040>
- Sharma, A., & Verbeke, W. J. M. I. (2020). Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). *Frontiers in Big Data*, 3. <https://www.frontiersin.org/articles/10.3389/fdata.2020.00015>
- Sheu, Y.-H. (2020). Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry*, 11, 551299. <https://doi.org/10.3389/fpsy.2020.551299>
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589–1604. <https://doi.org/10.1109/JBHI.2017.2767063>
- Shorter, E. (1997). *A history of psychiatry: From the era of the asylum to the age of Prozac* (pp. xii, 436). John Wiley & Sons.
- Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Jim Zheng, W., & Roberts, K. (2021). Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *Journal of Biomedical Informatics*, 115, 103671. <https://doi.org/10.1016/j.jbi.2020.103671>

- Stanghellini, G., & Broome, M. R. (2014). Psychopathology as the basic science of psychiatry. *The British Journal of Psychiatry: The Journal of Mental Science*, 205(3), 169–170. <https://doi.org/10.1192/bjp.bp.113.138974>
- Starke, G., Clercq, E. D., Borgwardt, S., & Elger, B. S. (2021). Computing schizophrenia: Ethical challenges for machine learning in psychiatry. *Psychological Medicine*, 51(15), 2515–2521. <https://doi.org/10.1017/S0033291720001683>
- Stoyanov, D., & Maes, M. H. (2021). How to construct neuroscience-informed psychiatric classification? Towards nomothetic networks psychiatry. *World Journal of Psychiatry*, 11(1), 1–12. <https://doi.org/10.5498/wjp.v11.i1.1>
- Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, 10(1), Article 1. <https://doi.org/10.1038/s41398-020-0780-3>
- Suhara, Y., Xu, Y., & Pentland, A. ‘Sandy’. (2017). DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. *Proceedings of the 26th International Conference on World Wide Web*, 715–724. <https://doi.org/10.1145/3038912.3052676>
- Thornton, T. (2020). Psychiatry’s inchoate wish for a paradigm shift and the biopsychosocial model of mental illness. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 229–239). Oxford University Press. <https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697>
- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359–380. <https://proceedings.mlr.press/v106/tonekaboni19a.html>
- Trivedi, M. H. (2016). Right patient, right treatment, right time: Biosignatures and precision medicine in depression. *World Psychiatry*, 15(3), 237–238. <https://doi.org/10.1002/wps.20371>
- Tsou, J. Y. (2021). Philosophy of Psychiatry. *Elements in the Philosophy of Science*. <https://doi.org/10.1017/9781108588485>
- Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet. Psychiatry*, 3(2), 171–178. [https://doi.org/10.1016/S2215-0366\(15\)00505-2](https://doi.org/10.1016/S2215-0366(15)00505-2)
- Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, 14(3), 270–277. <https://doi.org/10.1002/wps.20238>
- Wampold, B. E., & Budge, S. L. (2012). The 2011 Leona Tyler Award Address: The Relationship—and Its Relationship to the Common and Specific Factors of Psychotherapy. *The Counseling Psychologist*, 40(4), 601–623. <https://doi.org/10.1177/0011000011432709>
- Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*, 2nd ed (pp. x, 323). Routledge/Taylor & Francis Group.
- Wampold, B. E., Imel, Z. E., Bhati, K. S., & Johnson-Jennings, M. D. (2007). Insight as a Common Factor. In *Insight in psychotherapy* (pp. 119–139). American Psychological Association. <https://doi.org/10.1037/11532-006>

Well Being Trust. (2020, December 16). *A Unified Vision for the Future of Mental Health, Addiction, and Well-Being in the United States*. Well Being Trust.
<https://wellbeingtrust.org/press-releases/ceos-from-14-top-mental-health-organizations-join-together/>

WHO. (2019, May 2). *Special initiative for mental health (2019–2023)*.
[https://www.who.int/publications-detail-redirect/special-initiative-for-mental-health-\(2019-2023\)](https://www.who.int/publications-detail-redirect/special-initiative-for-mental-health-(2019-2023))

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, 190(6), 997–1016. <https://doi.org/10.1007/s11229-011-0055-x>

Wolpert, M., Pote, I., & Sebastian, C. L. (2021). Identifying and integrating active ingredients for mental health. *The Lancet Psychiatry*, 8(9), 741–743.
[https://doi.org/10.1016/S2215-0366\(21\)00283-2](https://doi.org/10.1016/S2215-0366(21)00283-2)

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.
<https://doi.org/10.1038/nn.4244>

Zachar, P., & Kendler, K. S. (2017). The Philosophy of Nosology. *Annual Review of Clinical Psychology*, 13, 49–71. <https://doi.org/10.1146/annurev-clinpsy-032816-045020>

Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288.
<https://doi.org/10.1007/s13347-019-00382-7>

Chapter 7:

Conclusion

The explosive progress of artificial intelligence (AI), powered by the success of deep learning (DL), has transformed many complex ethical and philosophical issues surrounding AI from academic curios into matters of immediate practical importance (e.g., Agrawal et al., 2024; Dung, 2024; Placani, 2024; Schwitzgebel & Garza, 2020; Seger et al., 2023). Though the predictive and generative power of modern AI promises much, the opacity of DL-based systems limits the extent to which their behaviour can be understood and casts doubt on whether they should be trusted in high-stakes environments (Bereska & Gavves, 2024; Čartolovni et al., 2022; Rudin, 2019). These concerns are referred to collectively as the *black box problem* (BBP) (Castelvecchi, 2016; Zednik, 2021). In this dissertation, I have provided a novel conceptual framework which can be applied to analyse the BBP (see Chapter 1). Further, I have presented a construal of *understanding* tailored to playing the central role in this framework (see Chapter 2) and provided a philosophical analysis of two prominent explainability approaches (see Chapters 3 and 4). Finally, I have demonstrated how my framework can be applied to analyse specific manifestations of the BBP in different domains (see Chapters 5 and 6).

I will conclude the dissertation with some final reflections on how my contributions address the BBP. To that end, this conclusion will be organised as follows. Chapter 7.1 will reiterate the motivation for my contributions, explaining why the complex structure of the BBP demands analysis which balances breadth and depth. Chapter 7.2 will highlight the most important insights gleaned from the dissertation, detailing how they make progress on open problems in the research field. Finally, Chapter 7.3 will suggest directions for future research building on the foundations laid in this dissertation.

7.1 Addressing the Wickedness of the Black Box Problem

A central theme of the current work is that the problems posed by the opacity of DL-based systems do not have straightforward, technical answers. Instead, the BBP is a *wicked*

cluster of complex, interrelated problems which require interdisciplinary collaboration to address (Langer et al., 2021). Here, I characterise the wicked nature of the BBP and explain why it demands piecemeal contributions that mix high-level, abstract analyses with detailed, domain-specific ones.

The BBP is a *wicked* problem in the sense that it is unbounded, ill-structured, and interdependent (Mitroff, 2019; Rittel & Webber, 1973). Let me clarify. The BBP is unbounded because a diverse array of AI systems may plausibly be deployed across almost any area of human activity (Boon, 2020; Brooks, 2021; Brown et al., 2021). This limits the generalisability of insights gleaned from studying particular manifestations. The BBP is ill-structured because it transcends disciplinary boundaries, encompassing technical, social, economic, psychological, legal, and philosophical concerns (Cross & Ramsey, 2021; Langer et al., 2021). This limits the utility of applying pre-existing disciplinary frameworks. And the BBP is interdependent because it is composed of entangled subproblems which cannot be evaluated in isolation, but only in conjunction with possible solutions to other subproblems (Mitroff, 2019). This limits the value of isolated conceptual work.

The wickedness of the BBP renders a comprehensive analysis of the problem out of scope for any single work. Thus, I suggest, a beneficial approach is one which balances breadth and depth. Such an approach ought to combine a high-level view of the dependency structure of the overall problem with detailed analyses of particularly relevant conceptual components and domain-specific dynamics.¹²³ To that end, in this dissertation, I have offered contributions at multiple levels of abstraction, with the aim of achieving the desired balance between breadth and depth. The broadest major contribution is my model of the BBP, presented in the introduction (Chapter 1.3.2). This model, I claim, captures the essential structure of opacity-based problems. It also shows *why* the BBP is best addressed in a piecemeal manner. Below, I explain how the BBP

¹²³ The rationale for this approach is that the demand for coherence across different levels of abstraction generates mutually beneficial constraints. More concretely, low-level conceptual terms, such as *understanding* or *desideratum*, are constrained to be construed such that they fit into the overall model. At the same time, the model has its structure shaped by domain-specific analyses which reveal, e.g., which conceptual terms are required to understand particular BBP manifestations.

model does this and summarise its key features before moving on to specific contributions in Chapter 7.2.

My BBP model serves several important functions. First, the model contextualises the rest of my work. It does so by situating the contributions of the dissertation in a unified framework with a shared set of conceptual terms and simplifying assumptions. The specific contributions then fill in details about that model, clarifying how the components that make it up should be construed, and showing how it can be applied to analyse BBP manifestations. Second, by highlighting what needs to be specified about model components to analyse particular manifestations of the BBP, the model demonstrates why a piecemeal approach is required. For example, *stakeholders* are described in terms of their relation to the opaque AI system (their *role*) and their knowledge, motivation, and resources (their *profile*). And human-system interactions are characterised in terms of their directness, their continuity, and the control the stakeholder has over the system (and/or vice versa). Given the heterogeneity of values the relevant variables can take, and the interdependence that holds between those variables, the dynamics that emerge in specific BBP manifestations are too varied to be captured by any single analysis. Thus, a piecemeal approach is not only practical, but necessary for addressing the BBP in a sufficiently nuanced and specific manner. Third, the BBP model itself contains noteworthy structural features. Prominently, it positions *understanding* as the central epistemic construct. This accords with other pragmatic approaches to the BBP and coheres with the practical goals of explainable AI (XAI) research (Fleisher, 2022; Nyrup & Robinson, 2022; Páez, 2019).

Having justified the piecemeal nature of my dissertation and explained the contribution of my BBP model. I now turn to more specific contributions.

7.2 Understanding, Explainability Approaches, and Desiderata Satisfaction: Addressing the Research Questions

In this dissertation, I addressed three questions related to the BBP. These are RQ1: How should *understanding* be construed in the context of the BBP? RQ2: How effective are extant explainability approaches? And RQ3: To what extent can specific manifestations

of the BBP be solved? In this section, I will explicate how my work contributes to answering these questions, highlighting important lessons about the BBP that can be learned from my analyses.

7.2.1 Understanding

The first research question concerns understanding. Recent philosophical work has focused on developing a general account of understanding (Baumberger et al., 2016; Wilkenfeld, 2013, 2019) or an account of *scientific* understanding in particular (de Regt & Dieks, 2005; Lipton, 2008; Strevens, 2013). In contrast with these approaches, I believe that an analysis specifically tailored to the particular demands of the BBP is required. In this context, the central question becomes: how *should* understanding be construed, such that it can play the role required of it by the BBP model? To that end, in *Conceptualising Understanding in Explainable AI (XAI): An Abilities-Based Approach* (CU), my co-authors and I sought to provide a practically valuable conceptualisation of understanding specifically designed for furthering the research agenda of XAI. The main research contributions of this project are threefold. First, we made a positive case that an abilities-based view of understanding is well-suited to practical analyses of the BBP. Second, we articulated distinct clusters of abilities that can be used to operationalise, measure, and evaluate understanding of AI systems. Third, we validated our approach by showing how these abilities can be used to specify epistemic requirements for desiderata satisfaction. This pragmatic view of understanding provides the cornerstone for further specific research contributions.

7.2.2 Explainability approaches

The second research question concerns the efficacy and potential of current explainability approaches (e.g., Barredo Arrieta et al., 2020; Geiger et al., 2022; Karimi et al., 2022; Olah et al., 2018; Roberts & Yaida, 2022). This question can be divided into two parts: First, what approaches to explaining opaque AI systems are there? And second, to what extent can they address the problems caused by the opacity of AI systems? With respect to the first question, I offered a novel breakdown of the *landscape* of XAI approaches (see the *Explainability approaches* subsection of Chapter 1.3.1). My characterisation of this landscape differs from pre-existing surveys and taxonomies in its

breadth. In particular, neither *mechanistic interpretability* nor *theoretical explainability* are typically included in characterisations of XAI (e.g., Buchholz, 2023; Carvalho et al., 2019; Fleisher, 2022). I believe this is mistaken. If the goal of XAI is to address the BBP, and reducing opacity is the means by which manifestations of the BPP can be solved, then it is crucial for theorists of the field to include all means of targeting opacity in when surveying the landscape.¹²⁴

With respect to the second question, viz., the efficacy of explainability approaches, my co-authors and I provided analyses of mechanistic interpretability (in *Explaining AI through Mechanistic Interpretability* (XAI-MI)) and ante-hoc explainability (in *Revisiting the Performance-Explainability Trade-Off* (RPET)). With respect to mechanistic interpretability, our main contributions were twofold. First, we characterised its distinctive character by casting it as a pattern discovery process involving functional decomposition of target systems (Bechtel & Richardson, 2010; Burnston, 2019; Kästner & Haueis, 2021). Second, we argued that its cost-benefit profile is distinctive, adding substance to the above claim that the approach is an important part of the XAI landscape. My co-authors and I also made two key contributions in our analysis of ante-hoc explainability. First, we argued that the preponderance of evidence suggests that performance-explainability trade-offs do obtain in *some* domains, cutting against claims that the ante-hoc explainability approach is universally superior. Second, we argued that adequate comparisons of explainability approaches must incorporate consideration of factors like development resources, data domain features, and particular risks and provided a framework for reasoning about such comparisons. These analyses assess explainability approaches by what they enable practitioners to *do*, thus explicitly building on the BBP model and the view of understanding presented in CU.

7.2.3 Desiderata satisfaction

The third research question addressed in my dissertation concerns the solubility of specific manifestations of the BBP. Addressing this question requires drilling into the details of particular domains to ascertain, first, whether the opacity of AI systems is

¹²⁴ I cannot be certain that my characterisation of the landscape is exhaustive. However, if there are more approaches, this only further justifies the need for a broader and more inclusive perspective.

preventing the satisfaction of some stakeholder-held desideratum, and second, whether explainability approaches have a plausible route to reducing that opacity. In *Don't Fear the Bogeyman: On Why There is no Prediction-Understanding Trade-Off for Deep Learning in Neuroscience* (DFB), my co-author and I considered the use of opaque artificial neural network (ANN) models in neuroscience (Cao & Yamins, 2024; Chirimuuta, 2020; Doerig et al., 2023; Richards et al., 2019). Though the opacity of these scientific models affects the pathway by which they can produce scientific understanding, we argued that there is good reason to think that a research strategy with ANNs at its heart *can* contribute to such understanding (construed as particular abilities). In *Understanding as a Bottleneck for the Data-Driven Approach to Psychiatric Science*, I considered the potential of data-driven approaches to improving clinical decision-making in psychiatry (Bickman, 2020; Bzdok & Meyer-Lindenberg, 2018; Fernandes et al., 2017). Here, the desideratum of interest is (improved) patient outcomes. According to my analysis, the data-driven approach has no clear way of targeting patient understanding, construed as possession of adaptive mental constructs (Dilthey, 1894/1977).

Beyond the object-level conclusions drawn in the papers above, several points of general importance for the BBP can be extracted from the analyses. The analysis in UB brings out the difficulty of *identifying* BBP manifestations. The BBP manifests when the opacity of an AI system prevents a stakeholder-held desideratum (in this case, patient outcomes). It is unclear whether the use of data-driven approaches in psychiatry satisfies this condition. Though it is plausible that the opacity of these models provides a barrier to deriving insights about mental illness that might aid practitioners and theorists in *their* understanding of psychiatric disorder, it is an open question whether such insights could be translated into improved clinical outcomes for patients. Further lessons about the BBP can be gleaned from DFB. These points are important for debates about AI's impact on scientific understanding. First, the ease with which something can be understood is an imperfect guide to the degree of understanding it can confer. This blocks the straightforward inference from the prima facie opacity of ANNs to their inability to induce scientific understanding. Second, models often confer understanding in combination with distributed complementary information. Taking these considerations into account

is crucial for reasoning about the value of AI in science. Again, these views build upon my model of the BBP and the ability-based approach to understanding.

7.2.4 Dissertation themes: Trade-Offs and breadth

Overall, I have shown how my conceptual framework for the BBP, in combination with an abilities-based approach to understanding and a measured assessment of the potential of explainability approaches, can be applied to deliver coherent analysis of BBP manifestations. In the course of the dissertation, two themes have been running in the background that warrant explicit articulation. These themes are, I claim, particularly important to consider for anybody trying to think clearly about the BBP.

The first theme is *trade-offs*. In addition to being an explicit subject of *Revisiting the Performance-Explainability Trade-Off* and *Don't Fear the Bogeyman*, trade-offs also feature prominently in our conclusions about mechanistic interpretability (viz., one *can* attain detailed explanations, but at great cost in terms of time and labour). Generally speaking, trade-offs are a ubiquitous feature of complex real-world problems. In BBP manifestations, it is rarely the case that any single course of action will simultaneously achieve the best possible outcome with respect to all desiderata that affected parties might hold. Thus, thinking clearly about trade-offs is critical to analysing wicked problems like the BBP (see also Chapter 7.3.2).

The second theme is the value of *breadth*. Several contributions in this dissertation involve adopting a wider perspective on a particular BBP-related subject in a way that materially affects downstream analysis. For one example, as argued above, breadth is important for characterising the XAI landscape. For another, RPET contends that trade-offs are best assessed broadly (viz., at the level of model development procedures, not model classes). For yet another, DFB claims that scientific understanding is best interpreted as distinct from (and broader than) the understanding of specific models. The general lesson, I take it, is that a broad perspective is often critical when reasoning about poorly understood problems (like the BBP). Construing key terms narrowly risks artificially severing possible values of upstream constructs (e.g., explainability approach) from the downstream ones we care about (viz., desiderata satisfaction). This is tantamount to characterising the problem space in a way that defines possible solutions

out of existence, and should be taken seriously as a methodological risk. For analysing wicked problems, I claim, breadth is a necessary complement to depth.

7.3 Looking Forward: Open Questions and Future Work

As I have emphasised, the BBP is a complicated and dynamic problem which will require further research to understand and continuous work to address. Before closing, I will identify four avenues for further areas of research that I believe are particularly interesting and important.

7.3.1 Apt descriptions for understanding AI systems

One topic of interest concerns how AI systems ought to be described. Since the release of ChatGPT in 2022, the use of chatbot-style AI systems based on the Transformer architecture has become commonplace among ordinary people (Draxler et al., 2023). These systems blur the boundaries between traditional tools and artificial agents, making it difficult to describe them using extant conceptual categories. This state of affairs raises numerous questions requiring further research. How should users of these systems conceptualise them so as to enable effective use? Which descriptions of these systems most reliably induce understanding of their capacities and limitations? And what underlying facts about these systems' behaviours and functional organisation justifies describing them in particular kinds of terms (e.g., in terms of cognitive attributes like understanding and creativity)? One approach to answering these questions involves building on the work presented in this dissertation. In particular, I believe focusing on which *abilities* are induced in recipients of descriptions is particularly promising. This will require interdisciplinary collaboration; cautious experimental designs are required to evaluate how varying descriptive information pertaining to chatbots and other AI systems affects users' understanding of them (Langer et al., 2022; Schlicker et al., 2021). Another angle on this question considers *scientific justification* for the attribution of descriptive terms (e.g., mental, cognitive, agential) to AI systems. Under the assumption that such attributions should depend on implementational (as opposed to strictly behavioural) criteria, high-quality mechanistic interpretability research is required to assess whether the functional organisation of LLMs and related systems fit the bill (this possibility was

mentioned in XAI-MI; see also Millière & Buckner, 2024). Ultimately, whichever descriptive policies one prefers for characterising novel AI systems, philosophical work will be required to justify them.

7.3.2 Trade-offs and the BBP

Another topic which deserves further scrutiny is trade-offs. This dissertation addressed several debates about trade-offs hypothesised to obtain in decision-making around opaque AI models. Though the domains under discussion in RPET and DFB differ in important ways,¹²⁵ some of the insights regarding the analysis of trade-offs applied to them both. In particular, my analysis suggests that the right level of abstraction to consider trade-offs is the level at which resource deployment and decision-making take place. Following this line of thought could lead to a more general analysis of trade-offs centred. In this dissertation, a general framework for thinking about trade-offs is sketched in RPET. There, the performance-explainability trade-off is described as a multi-objective optimisation problem centred around sequences of actions. However, due to space limitations, this analysis was underdeveloped. Future work could extend the conceptual framework, clarifying how the different pieces of conceptual machinery fit together and generalising it for the analysis of trade-offs across various domains in which the BBP manifests. This more detailed framework could then be applied to tackle outstanding questions in the philosophy of science about competing scientific values (Luk, 2018) and outstanding questions related to the BBP about, e.g., competing societal desiderata (Li & Li, 2024).

7.3.3 Mechanistic interpretability

The topic of mechanistic interpretability is another that remains significant. Since XAI-MI was conceived, much more technical work on mechanistic interpretability has been conducted (Bereska & Gavves, 2024; Marks et al., 2024; Rajamanoharan et al., 2024; Templeton et al., 2024). In XAI-MI, we took a broad perspective, focusing on outlining the distinctive features of MI from a philosophy of science perspective. Though we addressed some of the ongoing debates about MI in the worries and responses section of XAI-MI

¹²⁵ For example, scientific models typically have representational targets, while decision-support systems typically do not.

(Chapter 3.4), the brevity of our discussion was incommensurate with the richness of those debates. Thus, several issues are in need of more detailed treatment. One question concerns the goals of MI. When dealing with very large models, attaining complete and comprehensive explanations of their functional structure may be prohibitively costly and practically unattainable. Thus, practitioners must seek *mesoscopic* explanations that are abstract and simplified, while remaining (somewhat) faithful to how the system actually functions (Lindsay & Bau, 2023). How can these competing demands be balanced? What are the success conditions for such explanations? And how do those conditions vary as a function of the desiderata under consideration? Another question concerns alien concepts. It is widely agreed that *what* ANNs learn is not always straightforwardly interpretable by humans (Boge, 2022; Mann et al., 2023). A classic example is sensitivity to texture and noise patterns in early units of image processing ANNs (Geirhos et al., 2020). But do ANNs learn things that are *fundamentally* ungraspable for humans? Or can understanding and familiarity with these concepts be developed through interaction (as per model-induced understanding, see DFB)? A further question concerns the status of current hypotheses regarding the functional units of ANNs. Existing work isolates and investigates various computational structures including neurons, layers, connection weights, activation directions, circuits, and more besides (Bau et al., 2017; Cammarata et al., 2020; Geiger et al., 2022; Goh et al., 2021; Voss et al., 2021). Is a picture emerging about which levels of abstraction are most fruitful to study? And how generalisable are such findings across different models, architectures, and data domains? All of these questions are pressing and deserve philosophical attention.

7.3.4 Explainability approaches

A final topic for further research concerns explainability approaches. In the introduction to this dissertation, I distinguished between four explainability approaches and suggested that recognising their distinctive strengths and weaknesses is critical for understanding the XAI landscape.¹²⁶ To the best of my knowledge, no extant work directly

¹²⁶ An important caveat that bears repeating is that these approaches are not completely independent, either conceptually or sociologically. Particular research contributions may straddle the boundaries between approaches (Dreyer et al., 2024; Zhou et al., 2019), individual researchers may adopt different approaches at different times, and the approaches can be mutually beneficial in numerous ways.

addresses how these approaches interact with one another. Thus, spelling out the dynamic relations that hold between explainability approaches is an important subject for future work. Such work could address how theoretical explainability supports mechanistic interpretability (e.g., by clarifying how the mathematical substrate constrains mechanistic hypotheses), how method-centric explainability supports mechanistic interpretability (e.g., by providing methods that can be applied to test mechanistic hypotheses), how mechanistic interpretability supports ante-hoc explainability (e.g., by showing how subfunctions are implemented such that they can be explicitly designed), and more besides (initial ideas along these lines are sketched in XAI-MI). Future work could also clarify how explanatory information generated by the different explainability approaches can be integrated to develop a more comprehensive and holistic understanding of complex AI systems, building on the discussion in DFB concerning integrated understanding of biological neural systems.

Overall, there are numerous strands of research that must be conducted to improve our understanding of opaque AI systems. The selection above offers a tantalising taste of the complex dependency structures that must be navigated to address the BBP.

7.4 References

- Agrawal, A., Gans, J. S., & Goldfarb, A. (2024). Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy*, 33(2), 327–337. <https://doi.org/10.1111/jems.12521>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations. 3319–3327. <https://doi.org/10.1109/CVPR.2017.354>
- Baumberger, C., Beisbart, C., & Brun, G. (2016). What is Understanding? An Overview of Recent Debates in Epistemology and Philosophy of Science. In *Explaining Understanding*. Routledge.
- Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. The MIT Press; JSTOR. <https://doi.org/10.2307/j.ctt5hhhpX>

- Bereska, L., & Gavves, E. (2024, April 22). Mechanistic Interpretability for AI Safety—A Review. arXiv.Org. <https://arxiv.org/abs/2404.14082v3>
- Bickman, L. (2020). Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. *Administration and Policy in Mental Health*, 47(5), 795–843. <https://doi.org/10.1007/s10488-020-01065-8>
- Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1), 43–75. <https://doi.org/10.1007/s11023-021-09569-4>
- Boon, M. (2020). How Scientists Are Brought Back into Science—The Error of Empiricism. A Critical Reflection on Automated Science: Will Science Remain Human?, 43–65. https://doi.org/10.1007/978-3-030-25001-0_4
- Brooks, R. (2021). Artificial Intimacy: Virtual Friends, Digital Lovers, and Algorithmic Matchmakers. Columbia University Press. <https://doi.org/10.7312/broo20094>
- Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry*, 218(3), 131–134. <https://doi.org/10.1192/bjp.2019.245>
- Buchholz, O. (2023). A Means-End Account of Explainable Artificial Intelligence. *Synthese*, 202(2), 33. <https://doi.org/10.1007/s11229-023-04260-w>
- Burnston, D. C. (2019). Getting over Atomism: Functional Decomposition in Complex Neural Systems. *The British Journal for the Philosophy of Science*, axz039. <https://doi.org/10.1093/bjps/axz039>
- Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3. <https://doi.org/10.1016/j.bpsc.2017.11.007>
- Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, 5(3), 10.23915/distill.00024. <https://doi.org/10.23915/distill.00024>
- Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87, 101244. <https://doi.org/10.1016/j.cogsys.2024.101244>
- Čartolovni, A., Tomičić, A., & Lazić Mosler, E. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. *International Journal of Medical Informatics*, 161, 104738. <https://doi.org/10.1016/j.ijmedinf.2022.104738>
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8), 832. <https://doi.org/10.3390/electronics8080832>
- Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23. <https://doi.org/10.1038/538020a>
- Chirumuuta, M. (2020). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*. <https://doi.org/10.1007/s11229-020-02713-0>
- Cross, E. S., & Ramsey, R. (2021). Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. *Trends in Cognitive Sciences*, 25(3), 200–212. <https://doi.org/10.1016/j.tics.2020.11.009>

- de Regt, H. W., & Dieks, D. (2005). A Contextual Approach to Scientific Understanding. *Synthese*, 144(1), 137–170. <https://doi.org/10.1007/s11229-005-5000-4>
- Dilthey, W. (1977). Ideas Concerning a Descriptive and Analytic Psychology (1894) (R. M. Zaner, Trans.). In W. Dilthey (Ed.), *Descriptive Psychology and Historical Understanding* (pp. 21–120). Springer Netherlands. https://doi.org/10.1007/978-94-009-9658-8_2 (Original work published 1894)
- Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1–20. <https://doi.org/10.1038/s41583-023-00705-w>
- Draxler, F., Buschek, D., Tavast, M., Härmäläinen, P., Schmidt, A., Kulshrestha, J., & Welsch, R. (2023). Gender, Age, and Technology Education Influence the Adoption and Appropriation of LLMs (arXiv:2310.06556). arXiv. <https://doi.org/10.48550/arXiv.2310.06556>
- Dreyer, M., Purrelku, E., Vielhaben, J., Samek, W., & Lapuschkin, S. (2024). PURE: Turning Polysemantic Neurons Into Pure Features by Identifying Relevant Circuits (arXiv:2404.06453). arXiv. <https://doi.org/10.48550/arXiv.2404.06453>
- Dung, L. (2024). Understanding Artificial Agency. *The Philosophical Quarterly*, pqae010. <https://doi.org/10.1093/pq/pqae010>
- Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of ‘precision psychiatry’. *BMC Medicine*, 15(1), 80. <https://doi.org/10.1186/s12916-017-0849-x>
- Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, 19(4), 534–560. <https://doi.org/10.1017/epi.2022.39>
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing Causal Structure for Interpretable Neural Networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338. <https://proceedings.mlr.press/v162/geiger22a.html>
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), Article 11. <https://doi.org/10.1038/s42256-020-00257-z>
- Goh, G., † N. C., † C. V., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3), e30. <https://doi.org/10.23915/distill.00030>
- Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.*, 55(5), 95:1–95:29. <https://doi.org/10.1145/3527848>
- Kästner, L., & Haueis, P. (2021). Discovering Patterns: On the Norms of Mechanistic Inquiry. *Erkenntnis*, 86(6), 1635–1660. <https://doi.org/10.1007/s10670-019-00174-7>
- Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). “Look! It’s a Computer Program! It’s an Algorithm! It’s AI!”: Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems? *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–28. <https://doi.org/10.1145/3491102.3517527>

- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sasing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473. <https://doi.org/10.1016/j.artint.2021.103473>
- Li, J., & Li, G. (2024). The Triangular Trade-off between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey. *ACM Comput. Surv.* <https://doi.org/10.1145/3645088>
- Lindsay, G. W., & Bau, D. (2023). Testing methods of neural systems understanding. *Cognitive Systems Research*, 82, 101156. <https://doi.org/10.1016/j.cogsys.2023.101156>
- Lipton, P. (2008). Understanding Without Explanation. In H. W. D. Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific Understanding: Philosophical Perspectives* (pp. 43–63). University of Pittsburgh Press.
- Luk, R. W. P. (2018). To Explain or to Predict: Which One is Mandatory? *Foundations of Science*, 23(2), 411–414. <https://doi.org/10.1007/s10699-017-9533-z>
- Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023). Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy. 2023 IEEE 31st International Requirements Engineering Conference Workshops (REW), 337–342. <https://doi.org/10.1109/REW57809.2023.00063>
- Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models (arXiv:2403.19647). arXiv. <https://doi.org/10.48550/arXiv.2403.19647>
- Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models - Part II: The Way Forward (arXiv:2405.03207). arXiv. <http://arxiv.org/abs/2405.03207>
- Mitroff, I. I. (2019). Wicked Messes: The Pioneering Work of Horst Rittel and Russ Ackoff. In I. I. Mitroff (Ed.), *Technology Run Amok: Crisis Management in the Digital Age* (pp. 29–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-95741-8_4
- Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(1), 13. <https://doi.org/10.1007/s10676-022-09632-3>
- Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010. <https://doi.org/10.23915/distill.00010>
- Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>
- Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00419-4>
- Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., & Nanda, N. (2024). Improving Dictionary Learning with Gated Sparse Autoencoders (arXiv:2404.16014). arXiv. <https://doi.org/10.48550/arXiv.2404.16014>
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ...

- Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. <https://doi.org/10.1007/BF01405730>
- Roberts, D. A., & Yaida, S. (2022). *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press. <https://doi.org/10.1017/9781009023405>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Schlicker, N., Langer, M., Ötting, S. A., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up ‘black boxes’? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*, 122. <https://doi.org/10.1016/j.chb.2021.106837>
- Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780190905033.003.0017>
- Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple Meanings, Goals, and Methods. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 715–722. <https://doi.org/10.1145/3600211.3604693>
- Stevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, 44(3), 510–515. <https://doi.org/10.1016/j.shpsa.2012.12.005>
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>
- Voss, C., Cammarata, N., Goh, G., Petrov, M., Schubert, L., Egan, B., Lim, S., & Olah, C. (2021). Visualizing Weights. *Distill*, 6(2), 10.23915/distill.00024.007. <https://doi.org/10.23915/distill.00024.007>
- Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, 190(6), 997–1016. <https://doi.org/10.1007/s11229-011-0055-x>
- Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, 176(10), 2807–2831. <https://doi.org/10.1007/s11098-018-1152-1>
- Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288. <https://doi.org/10.1007/s13347-019-00382-7>
- Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of Deep Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep*

Chapter 7: Conclusion

Learning (pp. 243–252). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_12

Appendix A: Clarification of Specific Contributions

The papers making up chapters 2-5 of my dissertation were co-authored. Here, I provide more details on the specific contributions that I made to these joint works, with a focus on conceptual and structural contributions. Note that in addition to the specific contributions mentioned, I was involved in the writing and editing process for all parts of all papers.

Chapter 2: Conceptualising Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach

This paper was written with several other members of the Explainable Intelligent Systems research project. Timo Speith was the organisational lead for the paper. I was responsible for key aspects of the conceptualisation and structuring of the paper. First, I proposed that the structure of the paper should be based around delineating criteria a conceptualisation of understanding should satisfy and arguing for the abilities-based view on that basis. Second, I initially proposed the specific set of criteria. However, the way each criterion was ultimately construed in the final paper was a group decision. Third, I initially proposed the specific clusters of abilities we describe. Again, however, the way each ability cluster was ultimately construed in the final paper was a group decision. The published paper is the product of numerous re-writes and collective agreement on all content.

Chapter 3: Explaining AI Through Mechanistic Interpretability

This paper was co-authored with my supervisor, Prof. Lena Kästner. Prof. Kästner was the organisational lead for the paper. I was responsible for key aspects of the conceptualisation and structuring of the paper. First, I initially proposed the topic of the paper (viz., a philosophical analysis of mechanistic interpretability as an alternative approach to XAI). Second, I proposed an initial structure for the paper and wrote the first draft. Third, I identified the case study and proposed the set of worries and responses. The published paper is the product of numerous re-writes and collective agreement on all content.

Chapter 4: Revisiting the Performance-Explainability Trade-Off in Explainable Artificial Intelligence (XAI)

This paper was co-authored with Timo Speith, a member of the Explainable Intelligent Systems research project and colleague at Bayreuth University, and Maximilian Schlüter, a PhD student and computer scientist at TU Dortmund. I was responsible for key aspects of the conceptualisation and structuring of the paper. First, I initially proposed the topic of the paper (a critical evaluation of the explainability-performance trade-off). Second, I proposed the initial structure of the paper. Third, I devised the conceptual framework presented in the final section of the paper. The published paper is the product of collective agreement on all content.

Chapter 5: Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience

This paper was co-authored with my supervisor, Prof. Lena Kästner. I was responsible for key aspects of the conceptualisation and structuring of the paper. First, I suggested the topic for paper (a critical analysis of the claim that the use of deep learning models entails trading off understanding for predictive power, focusing on the argument presented in Chirumuuta (2021)). Second, I wrote the initial draft for the paper. The final paper is the product of numerous re-writes and collective agreement on all content.

Full Bibliography

- Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarencov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.
<https://doi.org/10.1016/j.inffus.2021.05.008>
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems*, 31.
<https://doi.org/10.48550/arXiv.1810.03292>
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2024). Artificial intelligence adoption and system-wide change. *Journal of Economics & Management Strategy*, 33(2), 327–337.
<https://doi.org/10.1111/jems.12521>
- Allahyari, H., & Lavesson, N. (2011). User-oriented assessment of classification model understandability. In Kofod-Petersen, A., Heintz, F., & Langseth, H. (Eds.), *Proceedings of the 11th Scandinavian Conference on Artificial Intelligence (SCAI 2011)*, Frontiers in Artificial Intelligence and Applications, Vol. 227. IOS Press, pp. 11–19,
<https://doi.org/10.3233/978-1-60750-754-3-11>
- Alon, U. (2006). An Introduction to Systems Biology: Design Principles of Biological Circuits. *Chapman and Hall/CRC*. <https://doi.org/10.1201/9781420011432>
- Alqaraawi, A., Schuessler, M., Weiß, P., et al. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In Paternò, F., Oliver, N., Conati, C., et al. (Eds.), *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI 2020)*. Association for Computing Machinery, pp. 275–285,
<https://doi.org/10.1145/3377325.3377519>
- Alvarez, M. (2009). How many kinds of reasons? *Philosophical Explorations*, 12(2), 181–193. <https://doi.org/10.1080/13869790902838514>
- American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders. *DSM Library*.
<https://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9780890425596>
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety (arXiv:1606.06565). *arXiv*.
<https://doi.org/10.48550/arXiv.1606.06565>
- Amorim, J. P., Abreu, P. H., Santos, J., Cortes, M., & Vila, V. (2023). Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. *Information Processing & Management*, 60(2), 103225.
<https://doi.org/10.1016/j.ipm.2022.103225>
- Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, 7, e479.
<https://doi.org/10.7717/peerj-cs.479>

- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Andersen, H. (2018). Complements, Not Competitors: Causal and Mathematical Explanations. *The British Journal for the Philosophy of Science*, 69(2), 485–508. <https://doi.org/10.1093/bjps/axw023>
- Anthony, W. A. (2000). A recovery-oriented service system: Setting some system level standards. *Psychiatric Rehabilitation Journal*, 24(2), 159–168. <https://doi.org/10.1037/h0095104>
- Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145(Pt B), 137–165. <https://doi.org/10.1016/j.neuroimage.2016.02.079>
- Arbib, M. (1961). Turing Machines, Finite Automata and Neural Nets. *Journal of the ACM*, 8(4), 467–475. <https://doi.org/10.1145/321088.321089>
- Artificial Intelligence Act (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf
- Baetu, T. M. (2011). Mechanism schemas and the relationship between biological theories. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences* (pp. 407–424). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199574131.003.0019>
- Baker, T. B., McFall, R. M., & Shoham, V. (2008). Current Status and Future Prospects of Clinical Psychology: Toward a Scientifically Principled Approach to Mental and Behavioral Health Care. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 9(2), 67–103. <https://doi.org/10.1111/j.1539-6053.2009.01036.x>
- Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., Jacob, A. P., Komeili, M., Konath, K., Kwon, M., Lerer, A., Lewis, M., Miller, A. H., Mitts, S., Renduchintala, A., Roller, S., ... Zijlstra, M. (2022). Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624), 1067–1074. <https://doi.org/10.1126/science.ade9097>
- Balaskas, A., Schueller, S. M., Cox, A. L., & Doherty, G. (2021). Ecological momentary interventions for mental health: A scoping review. *PLOS ONE*, 16(3), e0248152. <https://doi.org/10.1371/journal.pone.0248152>
- Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting Model Stitching to Compare Neural Representations. *Advances in Neural Information Processing Systems*, 34, 225–236. <https://doi.org/10.48550/arXiv.2106.07682>
- Baraniuk, R., Donoho, D., & Gavish, M. (2020). The science of deep learning. *Proceedings of the National Academy of Sciences*, 117(48), 30029–30032. <https://doi.org/10.1073/pnas.2020596117>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>

- Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439). <https://doi.org/10.1126/science.aav9436>
- Bastani, O., Kim, C., & Bastani, H. (2019). Interpreting Blackbox Models via Model Extraction. *arXiv*. <https://doi.org/10.48550/arXiv.1705.08504>
- Battaglia, P., Hamrick, J. B. C., Bapst, V., Sanchez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K., Nash, C., Langston, V. J., ... Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv*. <https://arxiv.org/pdf/1806.01261>
- Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network Dissection: Quantifying Interpretability of Deep Visual Representations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3319–3327. <https://doi.org/10.1109/CVPR.2017.354>
- Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology*, 35(1), 12. <https://doi.org/10.1007/s13347-022-00510-w>
- Baumberger, C. (2014). Types of understanding: Their nature and their relation to knowledge. *Conceptus*, 40(98), 67–88. <https://doi.org/10.1515/cpt-2014-0002>
- Baumberger, C. (2019). Explicating Objectual Understanding: Taking Degrees Seriously. *Journal for General Philosophy of Science*, 50(3), 367–388. <https://doi.org/10.1007/s10838-019-09474-6>
- Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). Routledge.
- Bechtel, W. (2009). Looking Down, Around, and Up: Mechanistic Explanation in Psychology. *Philosophical Psychology*, 22. <https://doi.org/10.1080/09515080903238948>
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–441. <https://doi.org/10.1016/j.shpsc.2005.03.010>
- Bechtel, W., & Richardson, R. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. MIT Press. <https://doi.org/10.7551/mitpress/8328.001.0001>
- Beck, A. T. (1985). Cognitive Therapy, Behavior Therapy, Psychoanalysis, and Pharmacotherapy. In M. J. Mahoney & A. Freeman (Eds.), *Cognition and Psychotherapy* (pp. 325–347). Springer US. https://doi.org/10.1007/978-1-4684-7562-3_14
- Beck, A. T., & Haigh, E. A. P. (2014). Advances in cognitive theory and therapy: The generic cognitive model. *Annual Review of Clinical Psychology*, 10, 1–24. <https://doi.org/10.1146/annurev-clinpsy-032813-153734>
- Beisbart, C. (2021). Opacity thought through: On the intransparency of computer simulations. *Synthese*, 199(3), 11643–11666. <https://doi.org/10.1007/s11229-021-03305-2>

- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849–15854.
<https://doi.org/10.1073/pnas.1903070116>
- Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It’s Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–266. <https://doi.org/10.1145/3531146.3533090>
- Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4.
<https://www.frontiersin.org/articles/10.3389/fdata.2021.688969>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
<https://doi.org/10.1145/3442188.3445922>
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798–1828. <https://doi.org/10.1109/tpami.2013.50>
- Bennett, D., Silverstein, S. M., & Niv, Y. (2019). The Two Cultures of Computational Psychiatry. *JAMA Psychiatry*, 76(6), 563–564.
<https://doi.org/10.1001/jamapsychiatry.2019.0231>
- Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety—A Review. *arXiv*. <https://arxiv.org/abs/2404.14082v3>
- Bickman, L. (2020). Improving Mental Health Services: A 50-Year Journey from Randomized Experiments to Artificial Intelligence and Precision Mental Health. *Administration and Policy in Mental Health*, 47(5), 795–843.
<https://doi.org/10.1007/s10488-020-01065-8>
- Bloom, B. S., Engelhart, M. D., Furst, E. J., et al. (1965). Taxonomy of educational objectives. *Handbook 1: Cognitive domain*. McKay.
- Boddington, P. (2023). Normative Ethical Theory and AI Ethics. In P. Boddington (Ed.), *AI Ethics: A Textbook* (pp. 229–276). Springer Nature. https://doi.org/10.1007/978-981-19-9382-4_6
- Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*, 32(1), 43–75. <https://doi.org/10.1007/s11023-021-09569-4>
- Boge, F. J. (2023). Functional Concept Proxies and the Actually Smart Hans Problem: What’s Special About Deep Neural Networks in Science. *Synthese*, 203(1), 16.
<https://doi.org/10.1007/s11229-023-04440-8>
- Bokulich, A. (2017). Models and Explanation. In L. Magnani & T. Bertolotti (Eds.), *Springer Handbook of Model-Based Science* (pp. 103–118). Springer International Publishing. https://doi.org/10.1007/978-3-319-30526-4_4
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). On the

Opportunities and Risks of Foundation Models. *arXiv*.

<https://doi.org/10.48550/arXiv.2108.07258>

Boon, M. (2020). How Scientists Are Brought Back into Science—The Error of Empiricism. In M. Bertolaso & F. Sterpetti (Eds.) *A Critical Reflection on Automated Science: Will Science Remain Human?*, 43–65. https://doi.org/10.1007/978-3-030-25001-0_4

Bordt, S., Finck, M., Raidl, E., et al. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In Isbell, C., Lazar, S., Oh, A., et al. (Eds.), *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAcT 2022)*. Association for Computing Machinery, pp. 891–905, <https://doi.org/10.1145/3531146.3533153>

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, 42, e2. <https://doi.org/10.1017/S0140525X17002266>

Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section D, Structural Biology*, 77(Pt 8), 982–991. <https://doi.org/10.1107/S2059798321007531>

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolphi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *The Behavioral and Brain Sciences*, 1–74. <https://doi.org/10.1017/S0140525X22002813>

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*.

Brigandt, I. (2011). Explanation in Biology: Reduction, Pluralism, and Explanatory Aims. *Science & Education*, 22(1), 69–91. <https://doi.org/10.1007/s11191-011-9350-7>

Bringsjord, S., & Govindarajulu, N. S. (2024). Artificial Intelligence. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/>

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv*. <http://arxiv.org/abs/2104.13478>

Brooks, R. (2021). *Artificial Intimacy: Virtual Friends, Digital Lovers, and Algorithmic Matchmakers*. Columbia University Press. <https://doi.org/10.7312/broo20094>

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1), 139–159. [https://doi.org/10.1016/0004-3702\(91\)90053-M](https://doi.org/10.1016/0004-3702(91)90053-M)

Brown, C., Story, G. W., Mourão-Miranda, J., & Baker, J. T. (2021). Will artificial intelligence eventually replace psychiatrists? *The British Journal of Psychiatry*, 218(3), 131–134. <https://doi.org/10.1192/bjp.2019.245>

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language

Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, 81(6), 1211–1241. <https://doi.org/10.1007/s10670-015-9791-5>

Brunotte, W., Chazette, L., Klös, V., & Speith, T. (2022). Quo Vadis, Explainability? – A Research Roadmap for Explainability Engineering. In V. Gervasi & A. Vogelsang (Eds.), *Requirements Engineering: Foundation for Software Quality* (Vol. 13216, pp. 26–32). Springer International Publishing. https://doi.org/10.1007/978-3-030-98464-9_3

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv*. <https://doi.org/10.48550/arXiv.2303.12712>

Buchholz, O. (2023). A Means-End Account of Explainable Artificial Intelligence. *Synthese*, 202(2), 33. <https://doi.org/10.1007/s11229-023-04260-w>

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, 195(12), 5339–5372. <https://doi.org/10.1007/s11229-018-01949-1>

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), e12625. <https://doi.org/10.1111/phc3.12625>

Budd, R., & Hughes, I. (2009). The Dodo Bird Verdict--controversial, inevitable and important: A commentary on 30 years of meta-analyses. *Clinical Psychology & Psychotherapy*, 16(6), 510–522. <https://doi.org/10.1002/cpp.648>

Burnston, D. C. (2016). Data graphs and mechanistic explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 57, 1–12. <https://doi.org/10.1016/j.shpsc.2016.01.002>

Burnston, D. C. (2021). Getting over Atomism: Functional Decomposition in Complex Neural Systems. *The British Journal for the Philosophy of Science*, 72(3), 743–772. <https://doi.org/10.1093/bjps/axz039>

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12. <https://doi.org/10.1177/2053951715622512>

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195301069.001.0001>

Buzsáki, G. (2020). The Brain–Cognitive Behavior Problem: A Retrospective. *eNeuro*, 7(4). <https://doi.org/10.1523/ENEURO.0069-20.2020>

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3. <https://doi.org/10.1016/j.bpsc.2017.11.007>

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, 5(3), 10.23915/distill.00024. <https://doi.org/10.23915/distill.00024>

Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve Circuits. *Distill*, 6(1), e00024.006. <https://doi.org/10.23915/distill.00024.006>

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. *arXiv*. <http://arxiv.org/abs/2104.01489>

- Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 87, 101244. <https://doi.org/10.1016/j.cogsys.2024.101244>
- Carnap, R. (1962). *Logical foundations of probability*. University of Chicago Press.
- Carruthers, P. (1987). Conceptual pragmatism. *Synthese*, 73(2), 205–224. <https://doi.org/10.1007/BF00484740>
- Čartolovni, A., Tomičić, A., & Lazić Mosler, E. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. *International Journal of Medical Informatics*, 161, 104738. <https://doi.org/10.1016/j.ijmedinf.2022.104738>
- Cartwright, N. (1994). *Nature's Capacities and Their Measurement* (1st ed.). Oxford University Press. <https://doi.org/10.1093/0198235070.001.0001>
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 1–34. <https://doi.org/10.3390/electronics8080832>
- Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23. <https://doi.org/10.1038/538020a>
- Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, 9(1), Article 1. <https://doi.org/10.1038/s41398-019-0607-2>
- Chang, H. (2012). *Is Water H₂O? Evidence, Realism and Pluralism*. Boston Studies in the Philosophy and History of Science.
- Charles Leek, E., Leonardis, A., & Heinke, D. (2022). Deep neural networks and image classification in biological vision. *Vision Research*, 197, 108058. <https://doi.org/10.1016/j.visres.2022.108058>
- Chazette, L., & Schneider, K. (2020). Explainability as a non-functional requirement: Challenges and recommendations. *Requirements Engineering*, 25(4), 493–514. <https://doi.org/10.1007/s00766-020-00333-1>
- Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. *2021 IEEE 29th International Requirements Engineering Conference (RE)*, 197–208. <https://doi.org/10.1109/RE51729.2021.00025>
- Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 20(2), 154–170. <https://doi.org/10.1002/wps.20882>
- Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet. Psychiatry*, 3(3), 243–250. [https://doi.org/10.1016/S2215-0366\(15\)00471-X](https://doi.org/10.1016/S2215-0366(15)00471-X)
- Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., & Cartocci, G. (2022). Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings.

- Computers in Human Behavior*, 137, 107406.
<https://doi.org/10.1016/j.chb.2022.107406>
- Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199(1), 767–790. <https://doi.org/10.1007/s11229-020-02713-0>
- Chollet, F. (2019). On the Measure of Intelligence. *arXiv*.
<https://doi.org/10.48550/arXiv.1911.01547>
- Chollet, F. (2021). *Deep Learning with Python*, Second Edition. Simon and Schuster.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2023). PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240), 1–113.
- Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science* (New York, N.Y.), 242(4879), 741–745. <https://doi.org/10.1126/science.3055294>
- Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, 23(4), 305–317. <https://doi.org/10.1016/j.tics.2019.01.009>
- Colaco, D. (2019). *An Investigation of Scientific Phenomena* [PhD Thesis]. University of Pittsburgh.
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. *arXiv*.
<https://doi.org/10.48550/arXiv.2304.14997>
- Coulombe, S., Radziszewski, S., Meunier, S., Provencher, H., Hudon, C., Roberge, P., Provencher, M. D., & Houle, J. (2016). Profiles of Recovery from Mood and Anxiety Disorders: A Person-Centered Exploration of People’s Engagement in Self-Management. *Frontiers in Psychology*, 7.
<https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00584>
- Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, 68(1), 53–74.
- Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3), 355–376.
<https://doi.org/10.1007/s11229-006-9097-x>
- Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford University Press.
- Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. University of Chicago Press.
<https://press.uchicago.edu/ucp/books/book/chicago/l/bo16123713.html>
- Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*, 71(1), 287–319. <https://doi.org/10.1093/bjps/axy015>
- Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4), 568–589. <https://doi.org/10.1086/709729>
- Cross, E. S., & Ramsey, R. (2021). Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. *Trends in Cognitive Sciences*, 25(3), 200–212.
<https://doi.org/10.1016/j.tics.2020.11.009>

- Cuevas, C. D. las, Peñate, W., & Rivera, L. de. (2014). To what extent is treatment adherence of psychiatric patients influenced by their participation in shared decision making? *Patient Preference and Adherence*, 8, 1547–1553. <https://doi.org/10.2147/PPA.S73029>
- Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, 11(1), 126. <https://doi.org/10.1186/1741-7015-11-126>
- de Boer, N. S., de Bruin, L. C., Geurts, J. J. G., & Glas, G. (2021). The Network Theory of Psychiatric Disorders: A Critical Assessment of the Inclusion of Environmental Factors. *Frontiers in Psychology*, 12, 221. <https://doi.org/10.3389/fpsyg.2021.623970>
- de Regt, H. W. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, 76(5), 585–597. <https://doi.org/10.1086/605795>
- de Regt, H. W. (2015). Scientific understanding: Truth or dare? *Synthese*, 192(12), 3781–3797. <https://doi.org/10.1007/s11229-014-0538-7>
- de Regt, H. W. (2019). From Explanation to Understanding: Normativity Lost? *Journal for General Philosophy of Science*, 50(3), 327–343. <https://doi.org/10.1007/s10838-019-09477-3>
- de Regt, H. W., & Dieks, D. (2005). A Contextual Approach to Scientific Understanding. *Synthese*, 144(1), 137–170. <https://doi.org/10.1007/s11229-005-5000-4>
- De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns*, 3(6), 100489. <https://doi.org/10.1016/j.patter.2022.100489>
- Deck, L., Schoeffer, J., De-Arteaga, M., et al. (2024). A critical survey on fairness benefits of XAI. In F. Steibel, M. Young & R. Baeza-Yates (Eds.), *Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2024). Association for Computing Machinery. <http://arxiv.org/abs/2310.13007>
- Deck, L., Schomäcker, A., Speith, T., Schöffner, J., Kästner, L., & Kühl, N. (2024). Mapping the Potential of Explainable AI for Fairness Along the AI Lifecycle. *arXiv*. <https://doi.org/10.48550/arXiv.2404.18736>
- Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7), 1829–1850.
- Dembic, S. (2023). Mental disorder: An ability-based view. *Philosophy and the Mind Sciences*, 4. <https://doi.org/10.33735/phimisci.2023.9630>
- Dennett, D. C. (1991). Real Patterns. *Journal of Philosophy*, 88(1), 27–51. <https://doi.org/10.2307/2027085>
- Department of Health and Social Care. (2021, March 27). Mental health recovery plan backed by £500 million. *GOV.UK*. <https://www.gov.uk/government/news/mental-health-recovery-plan-backed-by-500-million>
- Dickinson, E. (1998). *The Poems of Emily Dickinson*. Harvard University Press. (Original work published 1890)
- Dierckx, L., Veroneze, R., & Nijssen, S. (2023). RL-Net: Interpretable Rule Learning with Neural Networks. In H. Kashima, T. Ide, & W.-C. Peng (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 95–107). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33374-3_8

- Dilthey, W. (1977). Ideas Concerning a Descriptive and Analytic Psychology (1894) (R. M. Zaner, Trans.). In W. Dilthey (Ed.), *Descriptive Psychology and Historical Understanding* (pp. 21–120). Springer Netherlands. https://doi.org/10.1007/978-94-009-9658-8_2 (Original work published 1894)
- Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1–20. <https://doi.org/10.1038/s41583-023-00705-w>
- Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32. <https://dl.acm.org/doi/10.5555/3454287.3455504>
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78–87. <https://doi.org/10.1145/2347736.2347755>
- Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. <https://doi.org/10.23919/MIPRO.2018.8400040>
- Doyle, Y., Egan, S., Graham, N., & Khalifa, K. (2019). Non-factive Understanding: A Statement and Defense. *Journal for General Philosophy of Science*, 50(3), 345–365. <https://doi.org/10.1007/s10838-019-09469-3>
- Draxler, F., Buschek, D., Tavast, M., Hämmäläinen, P., Schmidt, A., Kulshrestha, J., & Welsch, R. (2023). Gender, Age, and Technology Education Influence the Adoption and Appropriation of LLMs. *arXiv*. <https://doi.org/10.48550/arXiv.2310.06556>
- Dreyer, M., Pürelku, E., Vielhaben, J., Samek, W., & Lapuschkin, S. (2024). PURE: Turning Polysemantic Neurons Into Pure Features by Identifying Relevant Circuits. *arXiv*. <https://doi.org/10.48550/arXiv.2404.06453>
- Dung, L. (2024). Understanding Artificial Agency. *The Philosophical Quarterly*, pqae010. <https://doi.org/10.1093/pq/pqae010>
- Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics-2020-106820. <https://doi.org/10.1136/medethics-2020-106820>
- Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, 24(11), Article 11. <https://doi.org/10.1038/s41380-019-0365-9>
- Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, 14, 91–118. <https://doi.org/10.1146/annurev-clinpsy-032816-045037>
- Eden, A. H. (2007). Three Paradigms of Computer Science. *Minds and Machines*, 17(2), 135–167. <https://doi.org/10.1007/s11023-007-9060-8>
- Ehrgott, M. (2005). *Multicriteria Optimization*. Springer Science & Business Media.
- Eldan, R., & Shamir, O. (2016). The Power of Depth for Feedforward Neural Networks. *Conference on Learning Theory*, 907–940. <https://proceedings.mlr.press/v49/eldan16.html>

- Elgin, C. (2007). Understanding and the facts. *Philosophical Studies*, 132(1), 33–42. <https://doi.org/10.1007/s11098-006-9054-z>
- Elgin, C. (2017). True enough. MIT Press.
- Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. *arXiv*. <https://doi.org/10.48550/arXiv.2209.10652>
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., ... Olah, C. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread*.
- Eliot, D. L. B. (2021). The Need For Explainable AI (XAI) Is Especially Crucial In The Law (SSRN Scholarly Paper 3975778). *Social Science Research Network*. <https://doi.org/10.2139/ssrn.3975778>
- Elujide, I., Fashoto, S. G., Fashoto, B., Mbunge, E., Folorunso, S. O., & Olamijuwon, J. O. (2021). Application of deep and machine learning techniques for multi-label classification performance on psychotic disorder diseases. *Informatics in Medicine Unlocked*, 23, 100545. <https://doi.org/10.1016/j.imu.2021.100545>
- Enck, P., & Zipfel, S. (2019). Placebo Effects in Psychotherapy: A Framework. *Frontiers in Psychiatry*, 10. <https://www.frontiersin.org/articles/10.3389/fpsy.2019.00456>
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64. <https://doi.org/10.1518/001872095779049543>
- Engel, G. L. (1977). The need for a new medical model: A challenge for biomedicine. *Science (New York, N.Y.)*, 196(4286), 129–136. <https://doi.org/10.1126/science.847460>
- Erasmus, A., Brunet, T. D. P., & Fisher, E. (2021). What is Interpretability? *Philosophy & Technology*, 34(4), 833–862. <https://doi.org/10.1007/s13347-020-00435-2>
- Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*, 77(1), 35–43. <https://doi.org/10.1001/jamapsychiatry.2019.2664>
- Ewbank, M. P., Cummins, R., Tablan, V., Catarino, A., Buchholz, S., & Blackwell, A. D. (2021). Understanding the relationship between patient language and outcomes in internet-enabled cognitive behavioural therapy: A deep learning approach to automatic coding of session transcripts. *Psychotherapy Research*, 31(3), 300–312. <https://doi.org/10.1080/10503307.2020.1788740>
- Facchini, A., & Termine, A. (2022). Towards a Taxonomy for the Opacity of AI Systems. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2021* (pp. 73–89). Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_7
- Faconti, G. P., & Massink, M. (2000). Continuity in human computer interaction. *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, 364. <https://doi.org/10.1145/633292.633511>

- Falcon, A. (2023). Aristotle on Causality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality/>
- Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning Hierarchical Features for Scene Labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1915–1929. <https://doi.org/10.1109/TPAMI.2012.231>
- Fawzi, A., Fawzi, H., & Fawzi, O. (2018). Adversarial vulnerability for any classifier. *Advances in Neural Information Processing Systems*, 31. <https://dl.acm.org/doi/10.5555/3326943.3327052>
- Faye, J. (2014). Not Just Why-questions. In J. Faye (Ed.), *The Nature of Scientific Thinking: On Interpretation, Explanation, and Understanding* (pp. 210–240). Palgrave Macmillan UK. https://doi.org/10.1057/9781137389831_9
- Faye, J. (2014). The Pragmatics of Explanation. In J. Faye (Ed.), *The Nature of Scientific Thinking: On Interpretation, Explanation, and Understanding* (pp. 183–209). Palgrave Macmillan UK. https://doi.org/10.1057/9781137389831_8
- Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, 23(7), 584–601. <https://doi.org/10.1016/j.tics.2019.03.009>
- Feest, U. (2017). Phenomena and Objects of Research in the Cognitive and Behavioral Sciences. *Philosophy of Science*, 84(5), 1165–1176.
- Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26(6), 3333–3361. <https://doi.org/10.1007/s11948-020-00276-4>
- Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of ‘precision psychiatry’. *BMC Medicine*, 15(1), 80. <https://doi.org/10.1186/s12916-017-0849-x>
- Firestone, C. (2020). Performance vs. Competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43), 26562–26571. <https://doi.org/10.1073/pnas.1905334117>
- First, M. B., Rebell, T. J., Keeley, J. W., Bhargava, R., Dai, Y., Kulygina, M., Matsumoto, C., Robles, R., Stona, A.-C., & Reed, G. M. (2018). Do mental health professionals use diagnostic classifications the way we think they do? A global survey. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, 17(2), 187–195. <https://doi.org/10.1002/wps.20525>
- Fleisher, W. (2022). Understanding, Idealization, and Explainable AI. *Episteme*, 19(4), 534–560. <https://doi.org/10.1017/epi.2022.39>
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Frances, A. (2009). Whither DSM-V? *The British Journal of Psychiatry: The Journal of Mental Science*, 195(5), 391–392. <https://doi.org/10.1192/bjp.bp.109.073932>

- Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. *Psychological Science in the Public Interest*, 22(3), 110–161. <https://doi.org/10.1177/15291006211051956>
- Frank, J. D., & Frank, J. B. (1993). *Persuasion and Healing: A Comparative Study of Psychotherapy*. JHU Press.
- Frankle, J., & Carbin, M. J. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv*. <https://doi.org/10.48550/arXiv.1803.03635>
- Freiesleben, T. (2024). Artificial Neural Nets and the Representation of Human Concepts. *arXiv*. <https://doi.org/10.48550/arXiv.2312.05337>
- Freitas, A. A. (2014). Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1), 1–10. <https://doi.org/10.1145/2594473.2594475>
- Frigg, R., & Hartmann, S. (2020). Models in Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>
- Garcez, A. d'Avila, & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11), 12387–12406. <https://doi.org/10.1007/s10462-023-10448-w>
- Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal*, 33(4), 111–117. <https://doi.org/10.1215/07402775-3813015>
- García-Gutiérrez, M. S., Navarrete, F., Sala, F., Gasparyan, A., Austrich-Olivares, A., & Manzanares, J. (2020). Biomarkers in Psychiatry: Concept, Definition, Types and Relevance to the Clinical Reality. *Frontiers in Psychiatry*, 11, 432. <https://doi.org/10.3389/fpsyt.2020.00432>
- Gauld, C., Micoulaud-Franchi, J.-A., & Dumas, G. (2021). Comment on Starke et al.: ‘Computing schizophrenia: ethical challenges for machine learning in psychiatry’: from machine learning to student learning: pedagogical challenges for psychiatry. *Psychological Medicine*, 51(14), 2509–2511. <https://doi.org/10.1017/S0033291720003906>
- Geer, A. J. (2021). Learning earth system models from observations: Machine learning or data assimilation? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 20200089. <https://doi.org/10.1098/rsta.2020.0089>
- Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems*, 34, 9574–9586. <https://dl.acm.org/doi/10.5555/3540261.3540994>
- Geiger, A., Potts, C., & Icard, T. (2023). Causal Abstraction for Faithful Model Interpretation. *arXiv*. <https://doi.org/10.48550/arXiv.2301.04709>
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing Causal Structure for Interpretable Neural Networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338. <https://proceedings.mlr.press/v162/geiger22a.html>
- Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), Article 11. <https://doi.org/10.1038/s42256-020-00257-z>

- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bygh9j09KX>
- Gelman, A., & Loken, E. (2019). The garden of forking paths: Why multiple comparisons can be a problem , even when there is no “ fishing expedition ” or “ p-hacking ” and the research hypothesis was posited ahead of time *. <https://www.semanticscholar.org/paper/The-garden-of-forking-paths-%3A-Why-multiple-can-be-a-Gelman-Loken/b63e25900013605c16f4ad74c636cfbd8e9a3e8e>
- Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science* (New York, N.Y.), 233(4771), 1416–1419. <https://doi.org/10.1126/science.3749885>
- Ghaemi, S. N. (2007). Existence and pluralism: The rediscovery of Karl Jaspers. *Psychopathology*, 40(2), 75–82. <https://doi.org/10.1159/000098487>
- Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750. [https://doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
- Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, 18, 34–42. <https://doi.org/10.1016/j.cobeha.2017.07.003>
- Glaser, J. I., Benjamin, A. S., Farhoodi, R., & Kording, K. P. (2019). The roles of supervised machine learning in systems neuroscience. *Progress in Neurobiology*, 175, 126–137. <https://doi.org/10.1016/j.pneurobio.2019.01.008>
- Glennan, S. (2017). *The new mechanical philosophy* (First edition). Oxford University Press.
- Glock, H.-J., & Schmidt, E. (2021). Pluralism about practical reasons and reason explanations. *Philosophical Explorations*, 24(2), 119–136. <https://doi.org/10.1080/13869795.2021.1908578>
- Glover, J. (2020). Psychiatry, folk psychology, and the impact of neuroscience—A response to Steve Hyman’s Loebel Lectures. In J. Savulescu, R. Roache, W. Davies, & essor J. P. Loebel (Eds.), *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 301–319). Oxford University Press.
- Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3). <https://doi.org/10.23915/distill.00030>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
- Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution, and phenomenology of the theory formation system. In *Explanation and cognition* (pp. 299–323). The MIT Press.
- Gosiewska, A., Kozak, A., & Biecek, P. (2021). Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. *Decision Support Systems*, 150, 113556. <https://doi.org/10.1016/j.dss.2021.113556>

- Gould, I. C., Shepherd, A. M., Laurens, K. R., Cairns, M. J., Carr, V. J., & Green, M. J. (2014). Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach. *NeuroImage: Clinical*, 6, 229–236. <https://doi.org/10.1016/j.nicl.2014.09.009>
- Grimm, S. (2021). Understanding. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2021/entries/understanding/>
- Grimm, S. R. (2006). Is Understanding A Species Of Knowledge? *The British Journal for the Philosophy of Science*. <https://doi.org/10.1093/bjps/axl015>
- Grimm, S. R. (2011). Understanding. In S. Bernecker & D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 84–94). <https://doi.org/10.4324/9780203839065.ch9>
- Grimm, S. R. (Ed.). (2019). *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*. Oxford University Press. <https://doi.org/10.1093/oso/9780190860974.001.0001>
- Grinvald, A., & Hildesheim, R. (2004). VSDI: A new era in functional imaging of cortical dynamics. *Nature Reviews. Neuroscience*, 5(11), 874–885. <https://doi.org/10.1038/nrn1536>
- Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of Explainable Artificial Intelligence. In M. Sayed-Mouchaweh (Ed.), *Explainable AI Within the Digital Transformation and Cyber Physical Systems* (pp. 9–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-76409-8_2
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 93:1-93:42. <https://doi.org/10.1145/3236009>
- Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2), Article 2. <https://doi.org/10.1609/aimag.v40i2.2850>
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- Guo, Z., Wu, Y., Hartline, J. D., & Hullman, J. (2024). A Decision Theoretic Framework for Measuring AI Reliance. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 221–236. <https://doi.org/10.1145/3630106.3658901>
- Gurova, L. (2022). The Uses of Truth: Is There Room for Reconciliation of Factivist and Non-Factivist Accounts of Scientific Understanding? *International Studies in the Philosophy of Science*, 35(3–4), 211–221. <https://doi.org/10.1080/02698595.2022.2084329>
- Ha, S., & Jeong, H. (2021). Unraveling hidden interactions in complex systems with deep learning. *Scientific Reports*, 11(1), Article 1. <https://doi.org/10.1038/s41598-021-91878-w>
- Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. <https://doi.org/10.1145/3593013.3594067>

- Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4), 889–911. <https://doi.org/10.1093/bjps/axi148>
- Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly*, 58(3), 269–290. <https://doi.org/10.2307/48616060>
- Harmer, C. J., Duman, R. S., & Cowen, P. J. (2017). How do antidepressants work? New perspectives for refining future treatment approaches. *The Lancet. Psychiatry*, 4(5), 409–418. [https://doi.org/10.1016/S2215-0366\(17\)30015-9](https://doi.org/10.1016/S2215-0366(17)30015-9)
- Harmer, C. J., Goodwin, G. M., & Cowen, P. J. (2009). Why do antidepressants take so long to work? A cognitive neuropsychological model of antidepressant drug action. *The British Journal of Psychiatry*, 195(2), 102–108. <https://doi.org/10.1192/bjp.bp.108.051193>
- Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2), 245–258. <https://doi.org/10.1016/j.neuron.2017.06.011>
- Haueis, P., & Kästner, L. (2022). Mechanistic inquiry and scientific pursuit: The case of visual processing. *Studies in History and Philosophy of Science*, 93, 123–135. <https://doi.org/10.1016/j.shpsa.2022.03.007>
- Haugeland, J. (1998). *Having Thought: Essays in the Metaphysics of Mind*. Harvard University Press.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Hempel, C. (1965). Aspects of Scientific Explanation. In *Philosophy and Phenomenological Research* (p. 504). Free Press.
- Hempel, C., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15, 135–175.
- Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2018). Deep Anomaly Detection with Outlier Exposure. *arXiv*. <https://doi.org/10.48550/arXiv.1812.04606>
- Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. *arXiv*. <https://doi.org/10.48550/arXiv.2306.12001>
- Hengartner, M. P., & Lehmann, S. N. (2017). Why Psychiatric Research Must Abandon Traditional Diagnostic Classification and Adopt a Fully Dimensional Scope: Two Solutions to a Persistent Problem. *Frontiers in Psychiatry*, 8, 101. <https://doi.org/10.3389/fpsy.2017.00101>
- Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). Natural Language Descriptions of Deep Visual Features. *arXiv*. <https://doi.org/10.48550/arXiv.2201.11114>
- Herzog, C. (2022). On the Ethical and Epistemological Utility of Explicable AI in Medicine. *Philosophy & Technology*, 35(2), 50. <https://doi.org/10.1007/s13347-022-00546-y>
- Hey, T., Butler, K., Jackson, S., & Thiayagalingam, J. (2020). Machine learning and big scientific data. *Philosophical Transactions of the Royal Society A: Mathematical*,

Physical and Engineering Sciences, 378(2166), 20190054.

<https://doi.org/10.1098/rsta.2019.0054>

Hickman, L., Bosch, N., Ng, V., et al. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8), 1323–1351. <https://doi.org/10.1037/apl0000695>

Hieronymi, P. (2011). XIV—Reasons for Action. *Proceedings of the Aristotelian Society*, 111(3_pt_3), 407–427. <https://doi.org/10.1111/j.1467-9264.2011.00316.x>

Hills, A. (2016). Understanding Why. *Noûs*, 50(4), 661–688.

<https://doi.org/10.1111/nous.12092>

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6), 82–97. <https://doi.org/10.1109/MSP.2012.2205597>

Hitchcock, C. R. (1995). Discussion: Salmon on Explanatory Relevance. *Philosophy of Science*, 62(2), 304–320. <https://doi.org/10.1086/289858>

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Hochstein, E. (2016). One mechanism, many models: A distributed theory of mechanistic explanation. *Synthese*, 193(5), 1387–1407.

<https://doi.org/10.1007/s11229-015-0844-8>

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. *arXiv*. <https://arxiv.org/abs/1812.04608v2>

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2024). Training compute-optimal large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030.

Holmes, A. J., & Patrick, L. M. (2018). The Myth of Optimality in Clinical Neuroscience. *Trends in Cognitive Sciences*, 22(3), 241–257. <https://doi.org/10.1016/j.tics.2017.12.006>

Horn, R. L., & Weisz, J. R. (2020). Can Artificial Intelligence Improve Psychotherapy Research and Practice. *Administration and Policy in Mental Health*, 47(5), 852–855. <https://doi.org/10.1007/s10488-020-01056-9>

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)

Howick, J., Glasziou, P., & Aronson, J. K. (2010). Evidence-based mechanistic reasoning. *Journal of the Royal Society of Medicine*, 103(11), 433–441. <https://doi.org/10.1258/jrsm.2010.100146>

Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, 471, 61–71. <https://doi.org/10.1016/j.canlet.2019.12.007>

Hullman, J., Holtzman, A., & Gelman, A. (2023). Artificial Intelligence and Aesthetic Judgment. *arXiv*. <https://doi.org/10.48550/arXiv.2309.12338>

- Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626. <https://doi.org/10.1007/s11229-008-9435-2>
- Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In Boyd, D., & Morgenstern, J. H. (Eds.), *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*. Association for Computing Machinery, pp. 49–58, <https://doi.org/10.1145/3287560.3287600>
- Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1), 141–154. <https://doi.org/10.1016/j.dss.2010.12.003>
- Hyman, S. E., & McConnell, D. (2020). Mental illness: The collision of meaning with mechanism. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 263–289). Oxford University Press. <https://oxfordmedicine.com/view/10.1093/med/9780198789697.001.0001/med-9780198789697>
- Imrie, F., Davis, R., & van der Schaar, M. (2023). Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. *Nature Machine Intelligence*, 5(8), 824–829. <https://doi.org/10.1038/s42256-023-00698-2>
- Insel, T. R., Collins, P. Y., & Hyman, S. E. (2015). Darkness Invisible: The Hidden Global Costs of Mental Illness. *Foreign Affairs*, 94(1), 127–135.
- Ivakhnenko, A. G. (1971). Polynomial Theory of Complex Systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(4), 364–378. <https://doi.org/10.1109/TSMC.1971.4308320>
- Ivanov, I., & Schwartz, J. M. (2021). Why Psychotropic Drugs Don't Cure Mental Illness—But Should They? *Frontiers in Psychiatry*, 12. <https://www.frontiersin.org/article/10.3389/fpsy.2021.579566>
- Ivanova, A. A., Hewitt, J., & Zaslavsky, N. (2021). Probing artificial neural networks: Insights from neuroscience. *arXiv*. <https://doi.org/10.48550/arXiv.2104.08197>
- Jacob, K. S. (2015). Recovery Model of Mental Illness: A Complementary Approach to Psychiatric Care. *Indian Journal of Psychological Medicine*, 37(2), 117–119. <https://doi.org/10.4103/0253-7176.155605>
- Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4–37. <https://doi.org/10.1109/34.824819>
- Janlert, L.-E., & Stolterman, E. (2017). The Meaning of Interactivity—Some Proposals for Definitions and Measures. *Human–Computer Interaction*, 32(3), 103–138. <https://doi.org/10.1080/07370024.2016.1226139>
- Janssen, R. J., Mourão-Miranda, J., & Schnack, H. G. (2018). Making Individual Prognoses in Psychiatry Using Neuroimaging and Machine Learning. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, 3(9), 798–808. <https://doi.org/10.1016/j.bpsc.2018.04.004>
- Janvid, M. (2018). Getting a Grasp of the Grasping Involved in Understanding. *Acta Analytica*, 33(3), 371–383. <https://doi.org/10.1007/s12136-018-0348-5>

- Jaspers, K. (1997). *General Psychopathology* (J. Hoenig & M. Hamilton, Trans.). JHU Press. (Original work published 1959)
- Johansson, H., & Eklund, M. (2003). Patients' opinion on what constitutes good psychiatric care. *Scandinavian Journal of Caring Sciences*, 17(4), 339–346. <https://doi.org/10.1046/j.0283-9318.2003.00233.x>
- Joober, R., & Tabbane, K. (2019). From the neo-Kraepelinian framework to the new mechanical philosophy of psychiatry: Regaining common sense. *Journal of Psychiatry & Neuroscience : JPN*, 44(1), 3–7. <https://doi.org/10.1503/jpn.180240>
- Jugovac, M., & Jannach, D. (2017). Interacting with Recommender Systems: Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems*, 7(3), 10:1-10:46. <https://doi.org/10.1145/3001837>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), Article 7873. <https://doi.org/10.1038/s41586-021-03819-2>
- Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, 46(3), 240–254. <https://doi.org/10.1016/j.tins.2022.12.008>
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv*. <https://doi.org/10.48550/arXiv.2001.08361>
- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6), Article 6. <https://doi.org/10.1038/s41593-019-0392-5>
- Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2022). A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.*, 55(5), 95:1-95:29. <https://doi.org/10.1145/3527848>
- Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic Recourse: From Counterfactual Explanations to Interventions. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 353–362. <https://doi.org/10.1145/3442188.3445899>
- Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2022). Towards Causal Algorithmic Recourse. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop* (pp. 139–166). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_8
- Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, 68, 68–79. <https://doi.org/10.1016/j.shpsa.2018.01.011>
- Kästner, L., & Haueis, P. (2021). Discovering Patterns: On the Norms of Mechanistic Inquiry. *Erkenntnis*, 86(6), 1635–1660. <https://doi.org/10.1007/s10670-019-00174-7>
- Kästner, L., Langer, M., Lazar, V., Schomacker, A., Speith, T., & Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. *2021 IEEE 29th*

- International Requirements Engineering Conference Workshops (REW)*, 169–175.
<https://doi.org/10.1109/REW53955.2021.00031>
- Keil, F. (2006). Explanation and Understanding. *Annual Review of Psychology*, 57(1), 227–254. <https://doi.org/10.1146/annurev.psych.57.102904.190100>
- Keil, F. (2019). How Do Partial Understandings Work? (pp. 191–208). In Grimm, S. R. (Ed.), *Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology*. Oxford University Press.
<https://doi.org/10.1093/oso/9780190860974.003.0010>
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644.e16. <https://doi.org/10.1016/j.neuron.2018.03.044>
- Kelp, C. (2015). Understanding phenomena. *Synthese*, 192(12), 3799–3816.
- Kendler, K. S. (2012). Levels of explanation in psychiatric and substance use disorders: Implications for the development of an etiologically based nosology. *Molecular Psychiatry*, 17(1), 11–21. <https://doi.org/10.1038/mp.2011.70>
- Kendler, K. S. (2016). The nature of psychiatric disorders. *World Psychiatry*, 15(1), 5–12.
<https://doi.org/10.1002/wps.20292>
- Kendler, K. S., & Campbell, J. (2014). Expanding the domain of the understandable in psychiatric illness: An updating of the Jasperian framework of explanation and understanding. *Psychological Medicine*, 44(1), 1–7.
<https://doi.org/10.1017/S0033291712003030>
- Kendler, K. S., & Gyngell, C. (2020). Multilevel interactions and the dappled causal world of psychiatric disorders. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 25–45). Oxford University Press.
<https://doi.org/10.1093/med/9780198789697.003.0003>
- Keogh, E., & Mueen, A. (2017). Curse of Dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer US.
https://doi.org/10.1007/978-1-4899-7687-1_192
- Khalifa, K. (2013). Understanding, grasping and luck. *Episteme*, 10(1), 1–17.
<https://doi.org/10.1017/epi.2013.6>
- Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press. <https://doi.org/10.1017/9781108164276>
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *Proceedings of the 35th International Conference on Machine Learning*, 2668–2677. <https://proceedings.mlr.press/v80/kim18d.html>
- Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (Un)reliability of Saliency Methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 267–280). Springer International Publishing.
https://doi.org/10.1007/978-3-030-28954-6_14

- Kirsch, A. (2017). Explain to whom? Putting the User in the Center of Explainable AI. *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017*. <https://www.semanticscholar.org/paper/Explain-to-whom-Putting-the-User-in-the-Center-of-Kirsch/da43f222f009a587381a2aa8e145f5a45c26db9f>
- Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science*, 48(4), 507–531.
- Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*, 84, 46–56. <https://doi.org/10.1016/j.shpsa.2020.08.003>
- Knuth, D. E. (1974). Structured Programming with go to Statements. *ACM Computing Surveys*, 6(4), 261–301. <https://doi.org/10.1145/356635.356640>
- Kohár, M., & Krickel, B. (2021). Compare and Contrast: How to Assess the Completeness of Mechanistic Explanation. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience* (pp. 395–424). Springer International Publishing. https://doi.org/10.1007/978-3-030-54092-0_17
- Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement. *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368. <https://doi.org/10.1109/RE.2019.00046>
- Koppe, G., Meyer-Lindenberg, A., & Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, 46(1), Article 1. <https://doi.org/10.1038/s41386-020-0767-z>
- Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Illankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., ... Meisenzahl, E. (2021). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry*, 78(2), 1–15. <https://doi.org/10.1001/jamapsychiatry.2020.3604>
- Kozachkov, L., Tauber, J., Lundqvist, M., Brincat, S. L., Slotine, J.-J., & Miller, E. K. (2022). Robust and brain-like working memory through short-term synaptic plasticity. *PLOS Computational Biology*, 18(12), e1010776. <https://doi.org/10.1371/journal.pcbi.1010776>
- Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Krishnan, M. (2020). Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning. *Philosophy & Technology*, 33(3), 487–502. <https://doi.org/10.1007/s13347-019-00372-9>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25. <https://dl.acm.org/doi/10.1145/3065386>
- Kvanvig, J. (2009). The value of understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 95–111). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199231188.003.0005>

- Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511498909>
- Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1), 33–67. <https://doi.org/10.1007/s13194-012-0056-8>
- Lam, N. (2022). Explanations in AI as Claims of Tacit Knowledge. *Minds and Machines*, 32(1), 135–158. <https://doi.org/10.1007/s11023-021-09588-1>
- Lange, M. (2017). *Because without cause: Non-causal explanations in science and mathematics*. Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780190269487.001.0001>
- Langer, M., & König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review*, 33(1), 100881.
<https://doi.org/10.1016/j.hrmr.2021.100881>
- Langer, M., Baum, K., Hartmann, K., et al. (2021). Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In Yue, T., & Mirakhorli, M. (Eds.), *Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops (REW 2021)*. IEEE, pp. 164–168,
<https://doi.org/10.1109/REW53955.2021.00030>
- Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment*, 29(2), 154–169.
<https://doi.org/10.1111/ijsa.12325>
- Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). “Look! It’s a Computer Program! It’s an Algorithm! It’s AI!”: Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems? *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–28.
<https://doi.org/10.1145/3491102.3517527>
- Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology*, 36(5), 751–769. <https://doi.org/10.1007/s10869-020-09711-6>
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, 296, 103473.
<https://doi.org/10.1016/j.artint.2021.103473>
- Lapuschkin, S., Binder, A., Montavon, G., et al. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In Tuytelaars, T., Li, F. F., Bajcsy, R., et al. (Eds.), *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. IEEE, pp. 2912–2920, <https://doi.org/10.1109/CVPR.2016.318>
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), Article 1. <https://doi.org/10.1038/s41467-019-08987-4>

- Laska, K. M., Gurman, A. S., & Wampold, B. E. (2014). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy* (Chicago, Ill.), 51(4), 467–481. <https://doi.org/10.1037/a0034332>
- Lawler, I., & Sullivan, E. (2021). Model Explanation Versus Model-Induced Explanation. *Foundations of Science*, 26(4), 1049–1074. <https://doi.org/10.1007/s10699-020-09649-1>
- Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1), 98–113. <https://doi.org/10.1109/72.554195>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), Article 7553. <https://doi.org/10.1038/nature14539>
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>
- Leder, G., & Zawidzki, T. (2023). The skill of mental health: Towards a new theory of mental health and disorder. *Philosophy and the Mind Sciences*, 4. <https://doi.org/10.33735/phimisci.2023.9684>
- Lederer, J. (2021). Activation Functions in Artificial Neural Networks: A Systematic Overview. *arXiv*. <https://doi.org/10.48550/arXiv.2101.09957>
- Legg, S., & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*: 17–24. <https://dl.acm.org/doi/10.5555/1565455.1565458>
- Leichsenring, F., Steinert, C., Rabung, S., & Ioannidis, J. P. A. (2022). The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: An umbrella review and meta-analytic evaluation of recent meta-analyses. *World Psychiatry*, 21(1), 133–145. <https://doi.org/10.1002/wps.20941>
- Levy, A., & Bechtel, W. (2013). Abstraction and the Organization of Mechanisms. *Philosophy of Science*, 80(2), 241–261.
- Li, J., & Li, G. (2024). The Triangular Trade-off between Robustness, Accuracy and Fairness in Deep Neural Networks: A Survey. *ACM Comput. Surv.* <https://doi.org/10.1145/3645088>
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. *arXiv*. <https://doi.org/10.48550/arXiv.2210.13382>
- Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network? *arXiv*. <https://doi.org/10.48550/arXiv.1907.06374>
- Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544
- Lindsay, G. W., & Bau, D. (2023). Testing methods of neural systems understanding. *Cognitive Systems Research*, 82, 101156. <https://doi.org/10.1016/j.cogsys.2023.101156>
- Lipton, P. (2008). Understanding Without Explanation. In H. W. D. Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific Understanding: Philosophical Perspectives* (pp. 43–63). University of Pittsburgh Press.

- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue*, 16(3), 31–57. <https://doi.org/10.1145/3236386.3241340>
- Liu, H., & Motoda, H. (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Springer US. <https://doi.org/10.1007/978-1-4615-5689-3>
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464–470. <https://doi.org/10.1016/j.tics.2006.08.004>
- Lombrozo, T. (2016). Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences*, 20(10), 748–759. <https://doi.org/10.1016/j.tics.2016.08.001>
- Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99(2), 167–204. <https://doi.org/10.1016/j.cognition.2004.12.009>
- London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1), 15–21. <https://doi.org/10.1002/hast.973>
- Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106, 102301. <https://doi.org/10.1016/j.inffus.2024.102301>
- López-Rubio, E., & Ratti, E. (2021). Data science and molecular biology: Prediction and mechanistic explanation. *Synthese*, 198(4), 3131–3156. <https://doi.org/10.1007/s11229-019-02271-0>
- Luk, R. W. P. (2018). To Explain or to Predict: Which One is Mandatory? *Foundations of Science*, 23(2), 411–414. <https://doi.org/10.1007/s10699-017-9533-z>
- Maimon, O., & Rokach, L. (Eds.). (2005). *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag. <https://doi.org/10.1007/b107408>
- Mann, S., Crook, B., Kästner, L., Schomäcker, A., & Speith, T. (2023). Sources of Opacity in Computer Systems: Towards a Comprehensive Taxonomy. *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, 337–342. <https://doi.org/10.1109/REW57809.2023.00063>
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054. <https://doi.org/10.1073/pnas.1907367117>
- Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an Integration of Deep Learning and Neuroscience. *Frontiers in Computational Neuroscience*, 10. <https://www.frontiersin.org/article/10.3389/fncom.2016.00094>
- Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *WIREs Data Mining and Knowledge Discovery*, 13(3), e1493. <https://doi.org/10.1002/widm.1493>
- Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. MIT.

- Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. *arXiv*. <https://doi.org/10.48550/arXiv.2403.19647>
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655. <https://doi.org/10.1016/j.jbi.2020.103655>
- Matthewson, J., & Weisberg, M. (2009). The structure of tradeoffs in model building. *Synthese*, 170(1), 169–190. <https://doi.org/10.1007/s11229-008-9366-y>
- Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars* (1st ed.). Cambridge University Press. <https://doi.org/10.1017/9781107286184>
- McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in Translation: Contextualized Word Vectors. *Advances in Neural Information Processing Systems*, 30. <https://dl.acm.org/doi/10.5555/3295222.3295377>
- McCarthy, J. (1979). Ascribing Mental Qualities to Machines. In M. Ringle (Ed.), *Philosophical Perspectives in Artificial Intelligence*. Humanities Press.
- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence: August 31, 1955. *AI Magazine*, 27(4), 12–14. <https://doi.org/10.1609/aimag.v27i4.1904>
- McConnell, D. (2020). The proper place of subjectivity, meaning, and folk psychology in psychiatry. In *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 290–303). Oxford University Press. <https://doi.org/10.1093/med/9780198789697.003.0018>
- McConnell, D., & Snoek, A. (2018). The Importance of Self-Narration in Recovery From Addiction. *Philosophy, Psychiatry, and Psychology*, 25(3), 31–44. <https://doi.org/10.1353/ppp.2018.0022>
- McCoy, L. G., Nagaraj, S., Morgado, F., Harish, V., Das, S., & Celi, L. A. (2020). What do medical students actually need to know about artificial intelligence? *Npj Digital Medicine*, 3(1), Article 1. <https://doi.org/10.1038/s41746-020-0294-7>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/BF02478259>
- McGorry, P. D., & Nelson, B. (2019). Transdiagnostic psychiatry: Premature closure on a crucial pathway to clinical utility for psychiatric diagnosis. *World Psychiatry*, 18(3), 359–360. <https://doi.org/10.1002/wps.20679>
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2023). Locating and Editing Factual Associations in GPT. *arXiv*. <https://doi.org/10.48550/arXiv.2202.05262>
- Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation Studies in Artificial Neural Networks. *arXiv*. <https://doi.org/10.48550/arXiv.1901.08644>
- Middleton, H., & Moncrieff, J. (2019). Critical psychiatry: A brief overview. *BJPsych Advances*, 25(1), 47–54. <https://doi.org/10.1192/bja.2018.38>

- Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, 100(2), 463–475. <https://doi.org/10.1016/j.neuron.2018.09.023>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum. Or: How I learnt to stop worrying and love the social and behavioural sciences. In Aha, D. W., Darrell, T., Pazzani, M., et al. (Eds.), *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI 2017)*. IJCAI, pp. 36–42, arXiv:1712.00547
- Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models -- Part I: Continuity With Classic Debates. *arXiv*. <https://doi.org/10.48550/arXiv.2401.03910>
- Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models - Part II: The Way Forward. *arXiv*. <http://arxiv.org/abs/2405.03207>
- Miranda, L., Paul, R., Pütz, B., Koutsouleris, N., & Müller-Myhsok, B. (2021). Systematic Review of Functional MRI Applications for Psychiatric Disease Subtyping. *Frontiers in Psychiatry*, 12, 665536. <https://doi.org/10.3389/fpsyt.2021.665536>
- Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120>
- Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy*, 17(1), 55–70. <https://doi.org/10.1023/A:1012990030867>
- Mitchell, S. D. (2019). Perspectives, Representation, and Integration. In *Understanding Perspectivism* (pp. 178–193). Routledge. <https://doi.org/10.4324/9781315145198-11>
- Mitchell, T. M. (2013). *Machine learning*. McGraw-Hill.
- Mitroff, I. I. (2019). Wicked Messes: The Pioneering Work of Horst Rittel and Russ Ackoff. In I. I. Mitroff (Ed.), *Technology Run Amok: Crisis Management in the Digital Age* (pp. 29–38). Springer International Publishing. https://doi.org/10.1007/978-3-319-95741-8_4
- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. <https://doi.org/10.1145/3287560.3287574>
- Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic Theory of Working Memory. *Science*, 319(5869), 1543–1546. <https://doi.org/10.1126/science.1150769>
- Morgan, M. S., & Morrison, M. (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press.
- Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics*, 3(3), 869–877. <https://doi.org/10.1007/s43681-022-00209-w>
- Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2(4), 345–389. <https://doi.org/10.1023/A:1009744630224>

- Nagel, E. (1982). *The structure of science* (Reprinted). Routledge and Kegan Paul.
- Najafpour, Z., Fatemi, A., Goudarzi, Z., Goudarzi, R., Shayanfar, K., & Noorizadeh, F. (2021). Cost-effectiveness of neuroimaging technologies in management of psychiatric and insomnia disorders: A meta-analysis and prospective cost analysis. *Journal of Neuroradiology*, 48(5), 348–358. <https://doi.org/10.1016/j.neurad.2020.12.003>
- Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *arXiv*. <https://doi.org/10.48550/arXiv.2301.05217>
- Nelson, R. J. (1978). The Competence-Performance Distinction in Mental Philosophy. *Synthese*, 39(3), 337–381.
- Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3), 113–126. <https://doi.org/10.1145/360018.360022>
- Newman, M. (2017). An evidentialist account of explanatory understanding. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 190–211). Routledge.
- Nguyen, A., Yosinski, J., & Clune, J. (2019). Understanding Neural Networks via Feature Visualization: A Survey. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 55–76). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_4
- Nie, Z., Vairavan, S., Narayan, V. A., Ye, J., & Li, Q. S. (2018). Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PLOS ONE*, 13(6), e0197268. <https://doi.org/10.1371/journal.pone.0197268>
- Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI Act. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-023-01723-z>
- Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, 27(3–5), 393–444. <https://doi.org/10.1007/s11257-017-9195-0>
- Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, 24(1), 13. <https://doi.org/10.1007/s10676-022-09632-3>
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3), e00024.001. <https://doi.org/10.23915/distill.00024.001>
- Olah, C., Cammarata, N., Voss, C., Schubert, L., & Goh, G. (2020). Naturally Occurring Equivariance in Neural Networks. *Distill*, 5(12), e00024.004. <https://doi.org/10.23915/distill.00024.004>
- Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, 2(11), 10.23915/distill.00007. <https://doi.org/10.23915/distill.00007>
- Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, 3(3), 10.23915/distill.00010. <https://doi.org/10.23915/distill.00010>

- Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>
- Paulus, M. P. (2015). Pragmatism Instead of Mechanism: A Call for Impactful Biological Psychiatry. *JAMA Psychiatry*, 72(7), 631–632. <https://doi.org/10.1001/jamapsychiatry.2015.0497>
- Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251), 20220041. <https://doi.org/10.1098/rsta.2022.0041>
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2. ed). Cambridge Univ. Press.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Pelin, H., Ising, M., Stein, F., Meinert, S., Meller, T., Brosch, K., Winter, N. R., Krug, A., Leenings, R., Lemke, H., Nenadić, I., Heilmann-Heimbach, S., Forstner, A. J., Nöthen, M. M., Opel, N., Repple, J., Pfarr, J., Ringwald, K., Schmitt, S., ... Andlauer, T. F. M. (2021). Identification of transdiagnostic psychiatric disorder subtypes using unsupervised learning. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, 46(11), 1895–1905. <https://doi.org/10.1038/s41386-021-01051-0>
- Perolat, J., De Vylder, B., Hennes, D., Tarassov, E., Strub, F., de Boer, V., Muller, P., Connor, J. T., Burch, N., Anthony, T., McAleer, S., Elie, R., Cen, S. H., Wang, Z., Gruslys, A., Malysheva, A., Khan, M., Ozair, S., Timbers, F., ... Tuyls, K. (2022). Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623), 990–996. <https://doi.org/10.1126/science.add4679>
- Pessoa, L. (2023). The Entangled Brain. *Journal of Cognitive Neuroscience*, 35(3), 349–360. https://doi.org/10.1162/jocn_a_01908
- Peterson, B. S. (2019). Editorial: Common factors in the art of healing. *Journal of Child Psychology and Psychiatry*, 60(9), 927–929. <https://doi.org/10.1111/jcpp.13108>
- Piltaver, R., Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, 62, 333–346. <https://doi.org/10.1016/j.eswa.2016.06.009>
- Piltaver, R., Luštreka, M., Gams, M., & Martinčić-Ipšić, S. (2014). Comprehensibility of classification trees-Survey design. In *Proceedings of 17th International Multiconference Information Society (IS 2014)*. Information Society, pp. 70–73
- Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00419-4>
- Poldrack, R. A., & Yarkoni, T. (2016). From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annual Review of Psychology*, 67(1), 587–612. <https://doi.org/10.1146/annurev-psych-122414-033729>
- Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). ‘Artiphysiology’ reveals V4-like shape tuning in a deep network trained for image classification. *ELife*, 7, e38242. <https://doi.org/10.7554/eLife.38242>

- Potochnik, A. (2010). Levels of explanation reconceived. *Philosophy of Science*, 77(1), 59–72. <https://doi.org/10.1086/650208>
- Potochnik, A. (2016). Scientific Explanation: Putting Communication First. *Philosophy of Science*, 2014 Symposium Proceedings, 721–732.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. <https://doi.org/10.1145/3411764.3445315>
- Priebe, S., Conneely, M., McCabe, R., & Bird, V. (2020). What can clinicians do to improve outcomes across psychiatric treatments: A conceptual review of non-specific components. *Epidemiology and Psychiatric Sciences*, 29. <https://doi.org/10.1017/S2045796019000428>
- Prince, S. J. D. (2023). *Understanding deep learning*. The MIT Press.
- Pu, X., & Kay, M. (2018). The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Paper. *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, 37–45. <https://doi.org/10.1109/BELIV.2018.8634103>
- Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., & Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nature Reviews Neuroscience*, 22(8), Article 8. <https://doi.org/10.1038/s41583-021-00473-5>
- Quiñonero-Candela, J. (Ed.). (2009). *Dataset shift in machine learning*. MIT Press.
- Qureshi, M. N. I., Min, B., Jo, H. J., & Lee, B. (2016). Multiclass Classification for the Differential Diagnosis on the ADHD Subtypes Using Recursive Feature Elimination and Hierarchical Extreme Learning Machine: Structural MRI Study. *PLOS ONE*, 11(8), e0160697. <https://doi.org/10.1371/journal.pone.0160697>
- Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). On the expressive power of deep neural net-works. *arXiv*. <https://doi.org/10.48550/arXiv.1606.05336>
- Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., & Nanda, N. (2024). Improving Dictionary Learning with Gated Sparse Autoencoders. *arXiv*. <https://doi.org/10.48550/arXiv.2404.16014>
- Ramon, S., Healy, B., & Renouf, N. (2007). Recovery from Mental Illness as an Emergent Concept and Practice in Australia and the UK. *International Journal of Social Psychiatry*, 53(2), 108–122. <https://doi.org/10.1177/0020764006075018>
- Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. *arXiv*. <https://doi.org/10.48550/arXiv.2207.13243>
- Räz, T. (2023). Methods for identifying emergent concepts in deep neural networks. *Patterns*, 4(6), 100761. <https://doi.org/10.1016/j.patter.2023.100761>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). ‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11491>
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. <https://doi.org/10.1038/s41593-019-0520-2>
- Riggs, W. D. (2003). Intellectual virtue: Perspectives from ethics and epistemology. In M. DePaul & L. Zagzebski (Eds.), *Understanding 'virtue' and the virtue of understanding* (pp. 203–226). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199252732.003.0010>
- Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences*, 4(2), 155–169. <https://doi.org/10.1007/BF01405730>
- Roberts, D. A., & Yaida, S. (2022). *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*. Cambridge University Press. <https://doi.org/10.1017/9781009023405>
- Rosenberg, N. (1982). *Inside the Black Box: Technology and Economics*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511611940>
- Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry*, 6, 412–415. <https://doi.org/10.1111/j.1939-0025.1936.tb05248.x>
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1: Foundations* (pp. 318–362). MIT Press.
- Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Parallel distributed processing models of schemata and sequential thought processes. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2, pp. 216–271). MIT Press.
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th Edition). Pearson. <http://aima.cs.berkeley.edu/>
- Rutledge, R. B., Chekroud, A. M., & Huys, Q. J. (2019). Machine learning and big data in psychiatry: Toward clinical applications. *Current Opinion in Neurobiology*, 55, 152–159. <https://doi.org/10.1016/j.conb.2019.02.006>
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

- Salmon, W. C., Jeffrey, R. C., & Greeno, J. G. (1971). *Statistical Explanation and Statistical Relevance*. University of Pittsburgh Press.
<https://doi.org/10.2307/j.ctt6wrd9p>
- Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700, pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer International Publishing. <https://doi.org/10.1007/978-3-030-28954-6>
- Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, 3(3), 210–229.
<https://doi.org/10.1147/rd.33.0210>
- Sarkar, A., Vijaykeerthy, D., Sarkar, A., & Balasubramanian, V. N. (2022). A Framework for Learning Ante-hoc Explainable Models via Concepts. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10276–10285.
<https://doi.org/10.1109/CVPR52688.2022.01004>
- Sathyanarayana Rao, T. S., & Andrade, C. (2016). Classification of psychotropic drugs: Problems, solutions, and more problems. *Indian Journal of Psychiatry*, 58(2), 111–113.
<https://doi.org/10.4103/0019-5545.183771>
- Schleiden, S., Friedrich, O., Gerlek, S., Assadi, G., & Seifert, J. (2023). The concept of “interaction” in debates on human–machine interaction. *Humanities and Social Sciences Communications*, 10(1), 551. <https://doi.org/10.1057/s41599-023-02060-8>
- Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In Schneegass, S., Pfleging, B., & Kern, D. (Eds.), *Proceedings of Mensch und Computer (MuC 2021)*. *Association for Computing Machinery*, pp. 325–329. <https://doi.org/10.1145/3473856.3474018>
- Schlicker, N., Langer, M., Ötting, S. A., Baum, K., König, C. J., & Wallach, D. (2021). What to expect from opening up ‘black boxes’? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*, 122.
<https://doi.org/10.1016/j.chb.2021.106837>
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Schmidhuber, J. (2022). Annotated History of Modern AI and Deep Learning. *arXiv*.
<https://doi.org/10.48550/arXiv.2212.11279>
- Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3613904.3642621>
- Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3), 413–423. <https://doi.org/10.1016/j.neuron.2020.07.040>
- Schubert, L., Voss, C., Cammarata, N., Goh, G., & Olah, C. (2021). High-Low Frequency Detectors. *Distill*, 6(1), e00024.005. <https://doi.org/10.23915/distill.00024.005>

- Schwitzgebel, E., & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780190905033.003.0017>
- Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple Meanings, Goals, and Methods. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 715–722. <https://doi.org/10.1145/3600211.3604693>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2), 336–359. <https://doi.org/10.1007/s11263-019-01228-7>
- Semenova, L., Rudin, C., & Parr, R. (2022). On the Existence of Simpler Machine Learning Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1827–1858. <https://doi.org/10.1145/3531146.3533232>
- Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. *Humanities and Social Sciences Communications*, 7(1), 1–10. <https://doi.org/10.1057/s41599-020-00614-8>
- Setiya, K. (2003). Explaining Action. *The Philosophical Review*, 112(3), 339–393.
- Shagrir, O., & Bechtel, W. (2017). Marr’s Computational Level and Delineating Phenomena. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780199685509.003.0009>
- Shalizi, C. R. (2006). Methods and Techniques of Complex Systems Science: An Overview. In T. S. Deisboeck & J. Y. Kresh (Eds.), *Complex Systems Science in Biomedicine* (pp. 33–114). Springer US. https://doi.org/10.1007/978-0-387-33532-2_2
- Sharma, A., & Verbeke, W. J. M. I. (2020). Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081). *Frontiers in Big Data*, 3. <https://www.frontiersin.org/articles/10.3389/fdata.2020.00015>
- Sheu, Y.-H. (2020). Illuminating the Black Box: Interpreting Deep Neural Network Models for Psychiatric Research. *Frontiers in Psychiatry*, 11, 551299. <https://doi.org/10.3389/fpsy.2020.551299>
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589–1604. <https://doi.org/10.1109/JBHI.2017.2767063>
- Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. Interface agents. *Interactions*, 4(6), 42–61. <https://doi.org/10.1145/267505.267514>
- Shorter, E. (1997). *A history of psychiatry: From the era of the asylum to the age of Prozac*. John Wiley & Sons.
- Si, Y., Du, J., Li, Z., Jiang, X., Miller, T., Wang, F., Jim Zheng, W., & Roberts, K. (2021). Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. *Journal of Biomedical Informatics*, 115, 103671. <https://doi.org/10.1016/j.jbi.2020.103671>

- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587). <https://doi.org/10.1038/nature16961>
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359. <https://doi.org/10.1038/nature24270>
- Simon, H. A. (2008). *The sciences of the artificial* (3. ed.). MIT Press.
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv*. <https://doi.org/10.48550/arXiv.1312.6034>
- Sloman, S. A., & Rabb, N. (2016). Your Understanding Is My Understanding: Evidence for a Community of Knowledge. *Psychological Science*, 27(11), 1451–1460. <https://doi.org/10.1177/0956797616662271>
- Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. <https://doi.org/10.1145/3351095.3372870>
- Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. <https://doi.org/10.1145/3531146.3534639>
- Speith, T. (2022). How to Evaluate Explainability? - A Case for Three Criteria. *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, 92–97. <https://doi.org/10.1109/REW56159.2022.00024>
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for Simplicity: The All Convolutional Net. *arXiv*. <https://doi.org/10.48550/arXiv.1412.6806>
- Srećković, S., Berber, A., & Filipović, N. (2022). The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation. *Minds and Machines*, 32(1), 159–183. <https://doi.org/10.1007/s11023-021-09575-6>
- Stanghellini, G., & Broome, M. R. (2014). Psychopathology as the basic science of psychiatry. *The British Journal of Psychiatry: The Journal of Mental Science*, 205(3), 169–170. <https://doi.org/10.1192/bjp.bp.113.138974>
- Starke, G., Clercq, E. D., Borgwardt, S., & Elger, B. S. (2021). Computing schizophrenia: Ethical challenges for machine learning in psychiatry. *Psychological Medicine*, 51(15), 2515–2521. <https://doi.org/10.1017/S0033291720001683>
- Steimers, A., & Schneider, M. (2022). Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*, 19(6), 3641. <https://doi.org/10.3390/ijerph19063641>
- Stirling, A. (2007). Risk, precaution and science: Towards a more constructive policy debate. Talking point on the precautionary principle. *EMBO Reports*, 8(4), 309–315. <https://doi.org/10.1038/sj.embor.7400953>

- Stock, P., Cissé, M. (2018). ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In Ferrari, V., Hebert, M., Sminchisescu, C., et al. (Eds.), *Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Lecture Notes in Computer Science*, Vol. 11210. Springer International Publishing, pp. 504–519, https://doi.org/10.1007/978-3-030-01231-1_31
- Stoyanov, D., & Maes, M. H. (2021). How to construct neuroscience-informed psychiatric classification? Towards nomothetic networks psychiatry. *World Journal of Psychiatry*, 11(1), 1–12. <https://doi.org/10.5498/wjp.v11.i1.1>
- Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press. <https://doi.org/10.2307/j.ctv1dv0tnw>
- Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, 44(3), 510–515. <https://doi.org/10.1016/j.shpsa.2012.12.005>
- Su, C., Xu, Z., Pathak, J., & Wang, F. (2020). Deep learning in mental health outcome research: A scoping review. *Translational Psychiatry*, 10(1), Article 1. <https://doi.org/10.1038/s41398-020-0780-3>
- Suhara, Y., Xu, Y., & Pentland, A. ‘Sandy’. (2017). DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. *Proceedings of the 26th International Conference on World Wide Web*, 715–724. <https://doi.org/10.1145/3038912.3052676>
- Sullivan, E. (2022). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1), 109–133. <https://doi.org/10.1093/bjps/axz035>
- Sussillo, D., Nuyujukian, P., Fan, J. M., Kao, J. C., Stavisky, S. D., Ryu, S., & Shenoy, K. (2012). A recurrent neural network for closed-loop intracortical brain-machine interface decoders. *Journal of Neural Engineering*, 9(2), 026027. <https://doi.org/10.1088/1741-2560/9/2/026027>
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3104–3112. <https://dl.acm.org/doi/10.5555/2969033.2969173>
- Sutton, R. (2019). *The bitter lesson*. <http://www.incompleteideas.net/Incldeas/BitterLesson.html>.
- Sutton, R. S., & Barto, A. (2020). *Reinforcement learning: An introduction* (Second edition). The MIT Press.
- Swanepoel, D. (2021). Does Artificial Intelligence Have Agency? In R. W. Clowes, K. Gärtner, & I. Hipólito (Eds.), *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts* (pp. 83–104). Springer International Publishing. https://doi.org/10.1007/978-3-030-72644-7_4
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html>

- Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovered the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. <https://doi.org/10.18653/v1/P19-1452>
- Thompson, J. A. F. (2021). Forms of explanation and understanding for neuroscience and artificial intelligence. *Journal of Neurophysiology*, 126(6), 1860–1874. <https://doi.org/10.1152/jn.00195.2021>
- Thórisson, K. R., Kremelberg, D., Steunebrink, B. R., et al. (2016). About understanding. In Steunebrink, B., Wang, P., & Goertzel, B. (Eds.), *Proceedings of the 9th International Conference on Artificial General Intelligence (AGI 2016), Lecture Notes in Computer Science*, Vol. 9782. Springer International Publishing, pp. 106–117, https://doi.org/10.1007/978-3-319-41649-6_11.
- Thornton, T. (2020). Psychiatry's inchoate wish for a paradigm shift and the biopsychosocial model of mental illness. In J. Savulescu, R. Roache, W. Davies, & J. P. Loebel (Eds.), *Psychiatry Reborn: Biopsychosocial psychiatry in modern medicine* (pp. 229–239). Oxford University Press. <https://doi.org/10.1093/med/9780198789697.003.0015>
- Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv*. <https://doi.org/10.48550/arXiv.1806.07552>
- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *Proceedings of the 4th Machine Learning for Healthcare Conference*, 359–380. <https://proceedings.mlr.press/v106/tonekaboni19a.html>
- Trivedi, M. H. (2016). Right patient, right treatment, right time: Biosignatures and precision medicine in depression. *World Psychiatry*, 15(3), 237–238. <https://doi.org/10.1002/wps.20371>
- Trültzsch-Wijnen, C. W. (2020). Theoretical Approach to the Concepts of Competence and Performance. In C. W. Trültzsch-Wijnen (Ed.), *Media Literacy and the Effect of Socialization* (pp. 93–126). Springer International Publishing. https://doi.org/10.1007/978-3-030-56360-8_4
- Tschantz, M. C. (2022). What is Proxy Discrimination? *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1993–2003. <https://doi.org/10.1145/3531146.3533242>
- Tsou, J. Y. (2021). Philosophy of Psychiatry. *Elements in the Philosophy of Science*. <https://doi.org/10.1017/9781108588485>
- Tullio, J., Dey, A. K., Chalecki, J., et al. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In Rosson, M. B., & Gilmore, D. J. (Eds.), *Proceedings of the 25th ACM Conference on Human Factors in Computing Systems (CHI 2007)*. Association for Computing Machinery, pp. 31–40, <https://doi.org/10.1145/1240624.1240630>.
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(October), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, 59(4), S251–S278.

- Van Camp, W. (2014). Explaining understanding (or understanding explanation). *European Journal for Philosophy of Science*, 4(1), 95–114. <https://doi.org/10.1007/s13194-013-0077-y>
- Van Fraassen, B. C. (1980). *The scientific image*. Clarendon press.
- Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2015). Goals affect the perceived quality of explanations. In Noelle DC, Dale R, Warlaumont AS, et al (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*, pp. 2469–2474, <https://cogsci.mindmodeling.org/2015/papers/0424/paper0424.pdf>
- Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review*, 24(5), 1436–1450. <https://doi.org/10.3758/s13423-017-1275-y>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. <https://dl.acm.org/doi/10.5555/3295222.3295349>
- Veit, W. (2020). Model Pluralism. *Philosophy of the Social Sciences*, 50(2), 91–114. <https://doi.org/10.1177/0048393119894897>
- Vigo, D., Thornicroft, G., & Atun, R. (2016). Estimating the true global burden of mental illness. *The Lancet. Psychiatry*, 3(2), 171–178. [https://doi.org/10.1016/S2215-0366\(15\)00505-2](https://doi.org/10.1016/S2215-0366(15)00505-2)
- Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3), 615–661. <https://doi.org/10.3390/make3030032>
- Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., ... Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350–354. <https://doi.org/10.1038/s41586-019-1724-z>
- Voss, C., Cammarata, N., Goh, G., Petrov, M., Schubert, L., Egan, B., Lim, S. K., & Olah, C. (2021). Visualizing Weights. *Distill*, 6(2), e00024.007. <https://doi.org/10.23915/distill.00024.007>
- Vredenburg, K. (2022). The right to explanation. *Journal of Political Philosophy*, 30(2), 209–229. <https://doi.org/10.1111/jopp.12262>
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. <https://doi.org/10.2139/ssrn.3063289>
- Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, 14(3), 270–277. <https://doi.org/10.1002/wps.20238>
- Wampold, B. E., & Budge, S. L. (2012). The 2011 Leona Tyler Award Address: The Relationship—and Its Relationship to the Common and Specific Factors of

- Psychotherapy. *The Counseling Psychologist*, 40(4), 601–623.
<https://doi.org/10.1177/0011000011432709>
- Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work*, (2nd ed). Routledge/Taylor & Francis Group.
- Wampold, B. E., Imel, Z. E., Bhati, K. S., & Johnson-Jennings, M. D. (2007). Insight as a Common Factor. In *Insight in psychotherapy* (pp. 119–139). *American Psychological Association*. <https://doi.org/10.1037/11532-006>
- Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small. *arXiv*.
<https://doi.org/10.48550/arXiv.2211.00593>
- Wang, P. (2019). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37. <https://doi.org/10.2478/jagi-2019-0002>
- Webb, A. R. (2003). *Statistical Pattern Recognition*. John Wiley & Sons.
- Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How Does LLM Safety Training Fail? *arXiv*. <https://doi.org/10.48550/arXiv.2307.02483>
- Well Being Trust. (2020, December 16). A Unified Vision for the Future of Mental Health, Addiction, and Well-Being in the United States. *Well Being Trust*.
<https://wellbeingtrust.org/press-releases/ceos-from-14-top-mental-health-organizations-join-together/>
- Wermter, S., & Sun, R. (2000). An Overview of Hybrid Neural Systems. In S. Wermter & R. Sun (Eds.), *Hybrid Neural Systems* (pp. 1–13). Springer.
https://doi.org/10.1007/10719871_1
- WHO. (2019, May 2). Special initiative for mental health (2019–2023). *WHO*.
[https://www.who.int/publications-detail-redirect/special-initiative-for-mental-health-\(2019-2023\)](https://www.who.int/publications-detail-redirect/special-initiative-for-mental-health-(2019-2023))
- Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, 190(6), 997–1016. <https://doi.org/10.1007/s11229-011-0055-x>
- Wilkenfeld, D. A. (2014). Functional explaining: A new approach to the philosophy of explanation. *Synthese*, 191(14), 3367–3391. <https://doi.org/10.1007/s11229-014-0452-z>
- Wilkenfeld, D. A. (2017). MUDdy understanding. *Synthese*, 194(4), 1273–1293.
<https://doi.org/10.1007/s11229-015-0992-x>
- Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, 176(10), 2807–2831. <https://doi.org/10.1007/s11098-018-1152-1>
- Wilkenfeld, D. A., & Hellmann, J. K. (2014). Understanding beyond grasping propositions: A discussion of chess and fish. *Studies in History and Philosophy of Science*, 48, 46–51. <https://doi.org/10.1016/j.shpsa.2014.09.003>
- Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., Cadena, S. A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., Sinz, F. H., & Tolias, A. S. (2023). Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. *bioRxiv*. <https://doi.org/10.1101/2023.05.12.540591>
- Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalisation: Evidence from category learning. *Cognitive Science*, 34(5), 776–806.
<https://doi.org/10.1111/j.1551-6709.2010.01113.x>

- Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In Ohlsson S, Catrambone R (Eds.), *Proceedings of the 32th Annual Meeting of the Cognitive Science Society*, pp. 2906–2911. <https://www.semanticscholar.org/paper/Why-does-explaining-help-learning-Insight-from-an-Williams-Lombrozo/c8b612ae56ea0e8e8f319c59909728e223676abb>
- Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination (1981). In L. Soler, E. Trizio, T. Nickles, & W. Wimsatt (Eds.), *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science* (pp. 61–87). Springer Netherlands. https://doi.org/10.1007/978-94-007-2759-5_2
- Wimsatt, W. C. (1994). The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy Supplementary Volume*, 20, 207–274. <https://doi.org/10.1080/00455091.1994.10717400>
- Winning, J., & Bechtel, W. (2018). Rethinking Causality in Biological and Neural Mechanisms: Constraints and Control. *Minds and Machines*, 28(2), 287–310. <https://doi.org/10.1007/s11023-018-9458-5>
- Winograd, T. (2006). Shifting viewpoints: Artificial intelligence and human–computer interaction. *Artificial Intelligence*, 170(18), 1256–1258. <https://doi.org/10.1016/j.artint.2006.10.011>
- Wolpert, M., Pote, I., & Sebastian, C. L. (2021). Identifying and integrating active ingredients for mental health. *The Lancet Psychiatry*, 8(9), 741–743. [https://doi.org/10.1016/S2215-0366\(21\)00283-2](https://doi.org/10.1016/S2215-0366(21)00283-2)
- Woodward, J. (2013). Mechanistic Explanation: Its Scope and Limits. *Proceedings of the Aristotelian Society*, Supplementary Volumes, 87, 39–65.
- Woodward, J. (2017). Explanation in Neurobiology: An Interventionist Perspective. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780199685509.003.0004>
- Woodward, J. F. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365. <https://doi.org/10.1038/nn.4244>
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624. <https://doi.org/10.1073/pnas.1403112111>
- Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking Bias-Variance Trade-off for Generalization of Neural Networks. *Proceedings of the 37th International Conference on Machine Learning*, 10767–10777. <https://dl.acm.org/doi/10.5555/3524938.3525936>
- Zachar, P., & Kendler, K. S. (2017). The Philosophy of Nosology. *Annual Review of Clinical Psychology*, 13, 49–71. <https://doi.org/10.1146/annurev-clinpsy-032816-045020>
- Zador, A. M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications*, 10(1), 3770. <https://doi.org/10.1038/s41467-019-11786-6>

- Zednik, C. (2015). Heuristics, Descriptions, and the Scope of Mechanistic Explanation. In P.-A. Braillard & C. Malaterre (Eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences* (pp. 295–318). Springer Netherlands. https://doi.org/10.1007/978-94-017-9822-8_13
- Zednik, C. (2021). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2), 265–288. <https://doi.org/10.1007/s13347-019-00382-7>
- Zednik, C., & Boelsen, H. (2022). Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines*, 32(1), 219–239. <https://doi.org/10.1007/s11023-021-09583-6>
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3), 107–115. <https://doi.org/10.1145/3446776>
- Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2019). Comparing the Interpretability of Deep Networks via Network Dissection. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 243–252). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_12
- Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5), Article 5. <https://doi.org/10.3390/electronics10050593>
- Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3), e2014196118. <https://doi.org/10.1073/pnas.2014196118>
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., ... Hendrycks, D. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv*. <https://doi.org/10.48550/arXiv.2310.01405>