Position: In Defence of Post-hoc Explainability

Nick Oh
socius
London, UK
nick.sh.oh@socius.org

Abstract

The widespread adoption of machine learning in scientific research has created a fundamental tension between model opacity and scientific understanding. Whilst some advocate for intrinsically interpretable models, we introduce Computational Interpretabilism (CI) as a philosophical framework for post-hoc interpretability in scientific AI. Drawing parallels with human expertise, where post-hoc rationalisation coexists with reliable performance, CI establishes that scientific knowledge emerges through structured model interpretation when properly bounded by empirical validation. Through mediated understanding and bounded factivity, we demonstrate how post-hoc methods achieve epistemic justification without requiring complete mechanical transparency, resolving tensions between model complexity and scientific comprehension.

1 Introduction

The increasing adoption of machine learning (ML) in scientific research has created a fundamental tension between the opacity of complex ML models and the need for scientific understanding. While ML models achieve unprecedented predictive performance across various scientific domains, from protein structure prediction to climate modeling, their complexity often renders them epistemically opaque – resistant to direct human understanding. This opacity presents a significant challenge to scientific practice, which traditionally relies on clear theoretical understanding and explanatory power.

The challenge has led to two dominant approaches in the ML community. One advocates for intrinsically interpretable models [Rudin, 2019], arguing that high-stakes (social) scientific applications demand transparent reasoning processes. The other develops post-hoc interpretability methods that attempt to explain already-trained complex models. This latter approach, while "practical", has faced mounting scepticism: recent empirical studies have revealed limitations in current post-hoc methods [Hooker et al., 2019, Adebayo et al., 2022, Bilodeau et al., 2024], raising deeper philosophical questions about their epistemic validity.

In response to this tension, particularly regarding the epistemic status of post-hoc interpretability methods in scientific AI applications, we introduce Computational Interpretabilism (CI), a philosophical stance that acknowledges that while complete, factive explanations of complex AI systems might be currently unattainable or pragmatically limiting, we can achieve epistemic value within bounded approximations. This argument rests on two key principles: (1) scientific knowledge can emerge through structured interpretation of model behaviour, even without direct access to the model's internal mechanisms; and (2) approximative explanations can provide epistemically justified scientific insights when when we carefully define their limitations and verify their reliability through empirical testing.

Interpretable AI: Past, Present and Future Workshop at NeurIPS 2024

2 On Interpretability

Human expertise has long been characterised by a distinctive gap between performance and "factive" explanation, particularly in domains where intuitive decision-making plays a central role [Dreyfus, 1972, Dreyfus and Dreyfus, 1986, Kahneman, 2003, Gobet and Chassy, 2008, 2009, Gobet, 2012]. Experts across domains routinely make effective decisions while their subsequent explanations often involve post-hoc rationalisation rather than complete, factive accounts of their decision-making processes [Bilalić et al., 2008a,b].

This gap between performance and explanation is not a flaw but a fundamental feature of expertise, reflecting how complex knowledge is encoded and deployed in human cognition. Our acceptance of this limitation in human expertise, contrasted with our skepticism toward similarly limited post-hoc explanations, raises important questions about interpretability in scientific AI systems. Can explanation and performance be epistemically decoupled without compromising the validity of either? What can the limitations of human expertise teach us about epistemically justifying post-hoc interpretability methods? These questions challenge us to reconsider not just the technical aspects of post-hoc models, but the fundamental nature of explanation and understanding in scientific practice.

Before exploring these implications further, it is essential to establish key definitions and assumptions to frame our discussion. We first conceptually fine-tune what we exactly mean by "interpretability".

Interpretability in AI systems is fundamentally pluralistic [Zednik, 2021], encompassing multiple distinct concepts and serving diverse stakeholder needs. As Lipton [2016] and Beisbart and Räz [2022] emphasise, interpretability is not a monolithic concept but rather reflects different expectations, questions, and explanatory virtues depending on the stakeholder and context.

Different stakeholders approach interpretability with distinct questions and needs. Computer scientists typically focus on understanding how inputs are mechanistically processed to produce outputs – what we might call mechanistic interpretability [Nanda et al., 2023]. In contrast, domain experts in scientific fields often seek to understand how model outputs inform or align with real-world phenomena – what we term phenomenological interpretability. While these perspectives are distinct, they are arguably reciprocal. As Ali et al. [2023] suggest, mechanistic interpretability often serves as a prerequisite for meaningful explanation, while the process of developing explanations can provide insights into the model's functioning.

For scientific applications, which are the focus of this paper, interpretability takes on additional dimensions beyond just understanding model mechanics. It encompasses our ability to understand and articulate how AI systems generate insights about natural phenomena in ways that advance scientific understanding. This broader conception is particularly important when considering post-hoc interpretability methods, whose criticisms often stem from evaluating them solely on their ability to faithfully explain a model's learned function ($h^*(X)$) through some interpretable approximation ($p^*(X)$). However, in scientific contexts, the key relationship is not just between $h^*(X)$ and $p^*(X)$, but how both relate to the underlying natural phenomenon $f(X)$ being studied. (A detailed visualisation and explanation of these functions and their relationships is provided in Appendix A.) This three-way relationship suggests that post-hoc methods can serve a vital epistemological role: bridging between model computations and scientific understanding of natural phenomena, even when they may not perfectly capture all details of model behaviour. To further contextualise our discussion, we establish several key assumptions:

1. *Accessibility of AI Systems*: When discussing post-hoc explanations, we concentrate on open and accessible black-box algorithms, rather than proprietary systems. The primary challenge in understanding these algorithms stems from their inherent complexity, rather than a complete lack of knowledge.

2. *Scientific AI Models*: Our analysis centres on supervised learning models designed to aid in science or knowledge discovery, such as predictive models in scientific research. We deliberately exclude discussion of interpretability in generative models, as they present distinct challenges beyond our current scope.

3. *Imperfect but Meaningful Approximations*: We assume post-hoc methods provide approximations with *imperfect but non-trivial fidelity* to the original model. By this, we explicitly exclude methods that perform no better than random attribution (e.g., Hooker et al. [2019], Bilodeau et al. [2024]), focusing instead on approaches that, while not perfect, demonstrably capture meaningful patterns in

model behaviour. While these approximations may not capture all nuances of model behaviour, they maintain sufficient accuracy to provide meaningful insights about both the model and the underlying phenomenon.

3 On Reliability and Justifiability

The parallel between human expertise and AI interpretability becomes particularly instructive when examining reliability and epistemic justifiability. Traditional epistemology distinguishes between two forms of justification [Pappas, 2005]: internalist (requiring accessible reasons) and externalist (focusing on reliable processes). This framework illuminates how both human expertise and post-hoc AI interpretability can achieve epistemic justification despite limited explicit articulation.

Human experts, too, primarily provide externalist justification. Their expertise stems from extensive experience and training, which develops incomprehensible (yet sophisticated) cognitive processes connecting features of their current experience with vast stores of domain-specific knowledge [Dreyfus and Dreyfus, 1984, 1986, Kahneman and Klein, 2009, Gobet, 2012]. Just as we trust our visual system without understanding its neural mechanisms, we accept expert judgment based on demonstrated reliability rather than complete explanation. Studies by Bilalić et al. [2008a,b] reveal that experts' attention patterns often diverge from their reported reasoning, suggesting their explicit explanations are post-hoc reconstructions rather than precise accounts of their decision processes.

This recognition of post-hoc rationalisation in human expertise reshapes our perspective on AI interpretability: if we accept human expert judgments despite their post-hoc nature, we might similarly justify post-hoc AI interpretability methods when they demonstrate reliable knowledge generation. The key is not perfect mechanistic transparency but rather reliable processes for generating and validating insights.

Moreover, this parallel suggests that requiring complete internalist justification (full explicit explanation) from AI systems may be not just practically challenging but philosophically unnecessary (e.g., also see Sullivan [2022]). While intrinsic interpretability aligns with internalist approaches by emphasising direct access to reasoning processes, the success of human expertise suggests this may be unnecessarily restrictive. Just as human expertise combines implicit pattern recognition with explicit domain knowledge, AI systems might achieve scientific validity through a combination of complex pattern recognition and post-hoc methods that reliably connect these patterns to domain knowledge.

This framework sets the stage for Computational Interpretabilism, which will show how post-hoc methods can achieve epistemic justification through bounded factivity and mediated understanding, even when their explanations are approximative. The key insight is that justification emerges not from complete mechanistic transparency but from reliable processes of knowledge generation and validation – a principle that applies equally to human expertise and AI systems.

4 Computational Interpretabilism

Computational Interpretabilism (CI) emerges as a philosophical framework that addresses one of the most pressing challenges in modern artificial intelligence: how to epistemically justify the use of post-hoc interpretability methods for scientific discovery. While machine learning models have demonstrated remarkable capabilities in analysing complex phenomena, their opacity raises fundamental questions about their role in generating valid scientific knowledge. CI provides a systematic approach to resolving this tension by establishing that post-hoc explanations can contribute legitimate scientific insights when they operate within specific epistemological boundaries [Beisbart and Räz, 2022]. The framework synthesises multiple philosophical perspectives, including Sullivan's link uncertainty, Andrews' theory-laden understanding, and Freiesleben et al.'s holistic representationality, to demonstrate how approximative explanations can bridge the gap between model complexity and scientific understanding. A broader treatment of philosophical perspectives on AI interpretability, including discussions beyond these frameworks, is provided in Appendix C.

4.1 Philosophical Foundations

The relationship between machine learning models and scientific understanding emerges as a complex interplay of multiple philosophical challenges and frameworks, particularly in justifying post-hoc interpretability methods for scientific inquiry. Freiesleben et al.'s distinction between elementwise representationality (ER) and holistic representationality (HR) provides an essential foundation[1] for understanding how modern ML models differ from traditional scientific models. This distinction gains deeper significance when viewed through multiple philosophical lenses: Andrews' theory-ladenness, Beisbart and Räz's factivity dilemma, Sullivan's link uncertainty, and fundamental principles from philosophy of science such as epistemic accessibility [Longino, 1990, Kitcher, 2001] and intersubjective verifiability[2] [Kuhn, 1997, Popper, 2005]. These perspectives together suggest that understanding emerges through mediated interpretation rather than direct access to model mechanics.

The shift from ER to HR that Freiesleben et al. propose can be understood as a practical response to the factivity dilemma – rather than attempting to maintain perfect fidelity at the component level (which would make models incomprehensible), HR acknowledges the necessity of holistic interpretation while providing rigorous methods for doing so through property descriptors. This approach aligns with Bilodeau et al.'s empirical findings that task-specific interpretability methods outperform general-purpose approaches, suggesting that effective scientific understanding requires targeted methods tailored to specific research contexts. However, as Andrews reminds us, these methods are inevitably theory-laden – the very choice of property descriptors and their implementation reflects our theoretical understanding and assumptions about both the phenomenon and the model. This theory-ladenness aligns with Douglas's (2009) recognition that scientific inquiry is inherently value-laden, requiring careful consideration of how interpretability methods mediate between model behaviour and human scientific understanding.

Sullivan's concept of link uncertainty provides a crucial bridge in understanding how post-hoc interpretability methods can be epistemically justified. These methods don't eliminate link uncertainty, but rather help manage it by mediating between model behaviour and phenomenal understanding through rigorous, testable connections. This connects to Popper's principle of falsifiability – interpretability methods must generate scrutinisable and potentially falsifiable claims about both model behaviour and phenomena. Beyond mere validation or falsification of existing knowledge, these methods can advance scientific understanding by revealing novel patterns and relationships that might not be apparent through traditional scientific approaches. The four-step framework proposed by Freiesleben et al. (formalisation, identification, estimation, and uncertainty quantification) can be seen as a systematic approach to reducing link uncertainty while acknowledging the theory-laden nature of scientific practice. Importantly, this framework demonstrates how post-hoc methods can be epistemically justified through their role in mediating between model behaviour, empirical validation, and scientific understanding.

The challenge of justifying post-hoc interpretability methods isn't just a technical problem but a fundamental epistemological challenge that requires careful attention to how scientific understanding emerges through mediated interpretation. Drawing on pragmatist philosophy [Putnam, 1995], we must recognise that interpretability's epistemic value lies in how it mediates between ML systems and human scientific understanding, facilitating oversight and integration with existing knowledge systems. Post-hoc interpretability methods emerge not just as technical tools for validation but as epistemological interfaces that actively participate in knowledge falsification and expansion, capable of generating new scientific insights while maintaining scientific rigour. They must make ML-generated knowledge accessible and comprehensible to the scientific community while enabling collaborative verification and critique, fundamentally shaping how we understand both ML systems and the phenomena they study.

Building on these philosophical foundations, CI establishes two key principles that together justify post-hoc interpretability methods in scientific ML. Mediated understanding reveals *how* scientific knowledge emerges through structured interactions between models, methods, and domain knowledge, while bounded factivity demonstrates *why* such mediated processes can be epistemically valid despite their approximative nature. Together, these principles provide a comprehensive philosophical

[1] The key difference is that CI sees interpretability methods not just as tools for extraction but as active participants in knowledge creation. This leads to the concept of "mediated understanding".

[2] That scientific claims should be verifiable by multiple observers.

framework that shows how post-hoc interpretability methods, when properly designed and validated, can contribute legitimately to scientific understanding.

4.1.1 Mediated Understanding

Scientific understanding through machine learning emerges not through direct model interpretation, but through a complex process of mediated interaction. The concept of "mediated understanding" in CI describes how scientific knowledge emerges through the structured interaction between four key elements: model behaviour, interpretability methods, domain knowledge, and empirical validation. This principle recognises that scientific understanding through ML is inherently mediated – direct access to model mechanics is not necessary for scientific insight [Sullivan, 2022, Beisbart and Räz, 2022]. Instead, understanding emerges through structured interpretation of model behaviour, and the relationship between model understanding and phenomenon understanding is reciprocal.

The epistemic validity of post-hoc methods in CI stems from their role as structured mediators in a bidirectional knowledge-creation process. In one direction (Model $\rightarrow$ Phenomenon), interpretability methods reveal patterns in model behaviour, which then *tentatively* suggest hypotheses about phenomena. These hypotheses, when tested empirically, provide new phenomenal understanding. In the other direction (Phenomenon $\rightarrow$ Model), domain knowledge guides the selection and refinement of interpretability methods, while empirical validation helps refine our interpretive approaches and identify relevant model behaviours for investigation. This bidirectional mediation provides epistemic justification because it ensures that interpretability methods are not merely describing model behaviour, but are actively participating in a cycle of hypothesis generation, empirical validation, and knowledge refinement – the very essence of scientific inquiry.

Consider a medical diagnosis ML system as an illustrative example [Sullivan, 2022]. Feature attribution methods serve as epistemic mediators by translating opaque model computations into testable hypotheses about biological mechanisms. When these model-derived insights are empirically validated against existing scientific evidence, they contribute to medical knowledge not despite their post-hoc nature, but precisely because their mediated interpretation enables systematic comparison between model behaviour and real-world phenomena. This demonstrates how CI's concept of mediated understanding resolves the apparent conflict between post-hoc interpretation and scientific validity – the very process of mediation, when properly structured and validated, becomes a legitimate source of scientific knowledge. A detailed examination of how mediated understanding operates differently across theory-rich versus theory-poor contexts is provided in Appendix B.

4.1.2 Bounded Factivity

Building on our discussion of mediated understanding, which shows how scientific knowledge emerges through structured interactions between models and methods, we now turn to another key concept that supports CI's defense of post-hoc interpretability: bounded factivity. This concept helps resolve fundamental tensions between the approximative nature of post-hoc methods and their epistemic value for scientific understanding.

In philosophy of science, the relationship between factivity and scientific understanding has been extensively debated. Traditional accounts often assume that genuine understanding requires strictly factual, true beliefs. However, recent work in philosophy of science has highlighted how idealisation and approximation play essential roles in scientific practice [Beisbart and Räz, 2022, Sullivan, 2022, Freiesleben et al., 2024]. Scientists routinely use simplified models that deliberately deviate from reality to gain understanding of complex phenomena. These non-factive elements, rather than being mere compromises, often prove essential for scientific progress.

This recognition of strategic simplification's role in science helps us reconceptualise the epistemic status of post-hoc interpretability methods. Rather than demanding complete factivity – perfect correspondence between interpretation and model mechanics – CI advocates for what we term "bounded factivity": truth within explicitly acknowledged limits and simplifications. Recent empirical work underscores the importance of this bounded approach. While Bilodeau et al. [2024] demonstrated that many popular post-hoc methods perform no better than random attribution, the authors also showed that carefully designed, task-specific approaches can provide reliable insights. By aligning interpretability methods with specific scientific goals [Freiesleben et al., 2024] and validating them through systematic empirical testing, we can achieve meaningful understanding within acknowledged bounds, just as traditional scientific models advance understanding despite their simplifications.

The concept of bounded factivity finds a natural parallel in Herbert Simon's bounded rationality [Simon, 1957, Wheeler, 2018] – both frameworks acknowledge inherent limitations while affirming the validity of strategic simplification. Just as bounded rationality accepts "satisficing" solutions under cognitive constraints, bounded factivity embraces carefully bounded approximations in interpretation. This pragmatic orientation helps justify post-hoc methods by showing how they can advance scientific understanding even without achieving perfect fidelity.

These parallels have important philosophical implications for how we justify post-hoc interpretability in scientific ML. Just as bounded rationality challenged idealised notions of human decision-making while preserving the validity of human judgment, bounded factivity challenges the assumption that perfect explanations are necessary for genuine scientific understanding while defending the epistemic value of post-hoc interpretations. It suggests that valid scientific understanding emerges through the careful management of trade-offs between accuracy and comprehensibility, realised through the process of mediated understanding where model behaviour, interpretive methods, and empirical validation interact cyclically. This mediated process, operating within explicitly acknowledged bounds, enables post-hoc methods to generate reliable scientific insights by systematically refining both our interpretation of models and our understanding of phenomena – precisely what well-designed post-hoc methods aim to achieve.

4.2 Re-assessing Criticisms of Post-hoc Models

4.2.1 Approximation and Fidelity

Rudin [2019] and Ghassemi et al. [2021] argue that post-hoc explanations are problematic due to their approximative nature, this critique necessitates careful examination of distinct but related concerns: the factivity of explanations and the nature of understanding in scientific practice. The dilemma presents itself thus: completely accurate explanations of complex ML models would merely duplicate their opacity, while simplified explanations necessarily introduce some degree of falsehood. This apparent tension can be productively addressed through the lens of non-factive understanding in science [Beisbart and Räz, 2022].

Interpretability exists on a spectrum, with increased epistemic value and practical utility correlating with higher degrees of interpretability. This approach aligns with the concept of verisimilitude, where approximations to truth, though imperfect, retain epistemic worth [Oddie, 2001]. Although post-hoc explanations lack performance guarantees and do not fully capture model behaviour, this limitation need not compromise their epistemological value if we maintain awareness of departure [Kvanvig, 2009] – conscious recognition of where and how our explanations diverge from ground truth. Many scientific and analytical tools rely on strategic idealisations that, despite their non-factive nature, provide valuable insights and practical utility. The key is maintaining empirical accountability through testable predictions, situating approximations within relevant theoretical frameworks, and providing clear scope conditions for the validity of interpretations.

This perspective aligns with CI's broader commitment to epistemic accessibility while acknowledging that accessibility often requires trade-offs with complete accuracy. Just as scientific models generally involve idealisations that technically violate factivity without compromising their utility for understanding, post-hoc explanations can provide genuine scientific insight even while containing strategic simplifications. This suggests that the key to maintaining scientific rigour lies not in perfect factivity, but in transparent acknowledgment of simplifications coupled with continuous refinement through empirical validation.

The relationship between model, interpretation, and reality can be illuminated through Korzybski's dictum, "The map is not the territory". Just as a map is a simplified representation of reality, both intrinsically interpretable models and post-hoc explanations are simplifications of the complex systems they represent. Accepting an intrinsically interpretable model as "understandable" and having some fidelity to the real world is philosophically analogous to accepting a post-hoc explanation that is "understandable" and has some fidelity to the original model. The fidelity between complex systems (real world or AI) and any model (intrinsic or post-hoc) is inherently imperfect, yet this imperfection does not negate their scientific value when properly bounded and validated.

4.2.2 Faithful Explanation and Confirmation Bias

Rudin [2019] identifies two potential pitfalls of post-hoc models – incomplete (local) explanations and unjustifiable explanations. The critique of local explanations underestimates their unique epistemological value in scientific practice. Rather than viewing local explanations as merely incomplete versions of global understanding, CI recognises them as distinct epistemic tools that offer granular insights into model behaviour. Through mediated understanding, these local insights can generate testable hypotheses about both model behaviour and phenomenal relationships, identify edge cases that reveal important patterns, and expose nuances that global explanations might miss. When properly bounded and validated, local explanations complement rather than compete with global understanding.

Regarding unjustifiable explanations, CI posits that even apparently problematic model behaviours – such as scientifically unsound judgments or confounding variables – can advance scientific understanding when properly interpreted. Through bounded factivity, we recognise that identifying flaws in model reasoning contributes valuable knowledge about both model limitations and phenomenal complexity. This aligns with how sciences historically progress through understanding both positive and negative results.

Ghassemi et al. [2021] raise a complementary concern about confirmation bias in interpreting post-hoc explanations, suggesting that humans might draw overconfident conclusions from potentially unreliable interpretations. This "interpretability gap" could potentially foster false confidence in the model's reliability or fairness. The limitations Ghassemi et al. describe are not unique to AI explanations but are inherent in complex judgements, whether human or artificial.

Human experts, like AI systems, can fall prey to confirmation bias, potentially leading to overconfidence in their interpretations or explanations. Following this reasoning to its logical conclusion, one might argue that we should be equally scepticalskeptical of human expert explanations as we are of AI-generated ones. Taken to an extreme, this line of thinking could lead to an argument for minimising reliance on expert explanations altogether, whether human or AI-generated. Instead, one might advocate for sole reliance on predefined, explicit "if-then" rules, aiming to eliminate the subjectivity and potential biases inherent in both human and AI interpretations. Yet, this conclusion overlooks the value of both human and AI-generated post-hoc explanations.

Rather than suggesting we should abandon post-hoc explanations in favor of purely rule-based approaches, CI advocates for their refinement and systematic validation. Just as sciences have developed methods for managing human cognitive biases while preserving the value of expert insight, we can develop approaches to post-hoc interpretation that acknowledge limitations while maximising epistemic value. This calls the field's attention to validation procedures and explicit acknowledgment of bounds.

5 Conclusion

Attempts to interpret AI models, particularly through post-hoc explanations, are analogous to human experts translating their intuitive impressions into deliberate explanations. Just as human experts often provide post-hoc rationalisations for their judgments, post-hoc explainability methods seek to make sense of the complex, compiled knowledge within AI. This process is inherently imperfect, as human experts also rarely access the full breadth of their internalised chunks and productions [Newell and Simon, 1972, Simon and Chase, 1973, Gobet and Clarkson, 2004]. However, this parallel reveals instructive differences in how we can approach AI interpretation. Unlike the black box of human cognition, AI systems permit systematic investigation of their internal states, allowing us to precisely specify the bounds of factivity and empirically verify our approximations. This accessibility enables a more rigorous approach to post-hoc interpretation than is possible with human expertise.

The epistemological framework of Computational Interpretabilism suggests that post-hoc interpretability methods serve a crucial mediating function in scientific ML, analogous to how expert explanations bridge specialised knowledge and broader understanding. When medical experts translate complex diagnoses for patients, they necessarily simplify their understanding into comprehensible explanations. Similarly, post-hoc interpretability methods transform opaque model computations into scientifically meaningful insights. The value lies not in perfect mechanical reproduction of the underlying process, but in reliable knowledge generation that can be validated against empirical evidence and integrated with domain expertise.

References

Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International conference on learning representations*, 2022.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99:101805, 2023.

Mel Andrews. The devil in the data: Machine learning & the theory-free ideal. 2023.

Claus Beisbart and Tim Räz. Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17(6):e12830, 2022.

Merim Bilalić, Peter McLeod, and Fernand Gobet. Why good thoughts block better ones: The mechanism of the pernicious einstellung (set) effect. *Cognition*, 108(3):652–661, 2008a.

Merim Bilalić, Peter McLeod, and Fernand Gobet. Inflexibility of experts—reality or myth? quantifying the einstellung effect in chess masters. *Cognitive psychology*, 56(2):73–102, 2008b.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

Heather E Douglas. *Science, policy, and the value-free ideal*. University of Pittsburgh Pre, 2009.

H.L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972. ISBN 9780060110826.

Hubert Dreyfus and Stuart E Dreyfus. *Mind over machine*. Simon and Schuster, 1986.

Hubert L Dreyfus and Stuart E Dreyfus. From socrates to expert systems: The limits of calculative rationality. *Technology in Society*, 6(3):217–233, 1984.

Juan M Durán and Nico Formanek. Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, 28:645–666, 2018.

Juan Manuel Durán. Machine learning, justification, and computational reliabilism. 2023.

Timo Freiesleben. Artificial neural nets and the representation of human concepts. *arXiv preprint arXiv:2312.05337*, 2023.

Timo Freiesleben, Gunnar König, Christoph Molnar, and Álvaro Tejero-Cantero. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. *Minds and Machines*, 34(3):32, 2024.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

Fernand Gobet. Concepts without intuition lose the game: commentary on montero and evans (2011). *Phenomenology and the Cognitive Sciences*, 11:237–250, 2012.

Fernand Gobet and Philippe Chassy. Towards an alternative to benner's theory of expert intuition in nursing: a discussion paper. *International journal of nursing studies*, 45(1):129–139, 2008.

Fernand Gobet and Philippe Chassy. Expertise and intuition: A tale of three theories. *Minds and Machines*, 19: 151–180, 2009.

Fernand Gobet and Gary Clarkson. Chunks in expert memory: Evidence for the magical number four... or is it two? *Memory*, 12(6):732–747, 2004.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Paul Humphreys. The philosophical novelty of computer simulation methods. *Synthese*, 169:615–626, 2009.

Noman Javed, Angelo Pirrone, Laura Bartlett, Peter Lane, and Fernand Gobet. Trust in cognitive models: Understandability and computational reliabilism. In *Proceedings of the AISB Convention 2023*, 2023.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.

Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: a failure to disagree. *American psychologist*, 64(6):515, 2009.

Philip Kitcher. *Science, truth, and democracy*. Oxford University Press, 2001.

Thomas S Kuhn. *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago, 1997.

Jonathan Kvanvig. Responses to critics. *Epistemic value*, pages 339–351, 2009.

Seth Lazar. Legitimacy, authority, and democratic. *Oxford Studies in Political Philosophy Volume 10*, page 28, 2024.

Zachary Lipton. The mythos of model interpretability. *Communications of the ACM*, 61:36 – 43, 2016.

Helen E Longino. *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton university press, 1990.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Allen Newell and Herbert Alexander Simon. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

Graham Oddie. Truthlikeness. *The Stanford Encyclopedia of Philosophy*, 2001.

George Pappas. Internalist vs. externalist conceptions of epistemic justification. 2005.

Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

Hilary Putnam. *Pragmatism: An Open Question*. Blackwell, 1995.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Herbert A. Simon and William G. Chase. Skill in chess. *American Scientist*, 61(4):394–403, 1973.

Herbert Alexander Simon. *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. New York: Wiley, 1957.

Emily Sullivan. Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 2022.

Juntao Tan and Yongfeng Zhang. Explainablefold: Understanding alphafold prediction with explainable ai. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2166–2176, 2023.

Kate Vredenburgh. The right to explanation. *Journal of Political Philosophy*, 30(2):209–229, 2022.

Gregory Wheeler. Bounded rationality. 2018.

Carlos Zednik. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & technology*, 34(2):265–288, 2021.

A Functional Relationships in Interpretable and Post-hoc AI Models

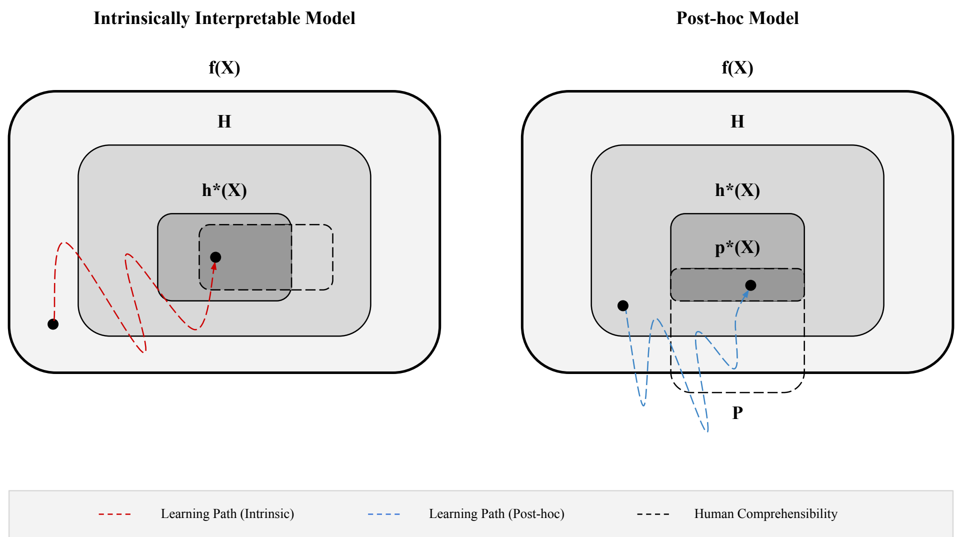

Figure 1: A comparative analysis of intrinsically interpretable and post-hoc explainable AI models.

This diagram illustrates the key functional relationships in both intrinsically interpretable and post-hoc explainable AI models within scientific contexts. The framework is grounded in three fundamental functions: the true phenomenon $f(X)$, the learned model $h^*(X)$, and for post-hoc methods, the interpretable approximation $p^*(X)$.

In both cases, we begin with a true function $f(X)$ representing the natural phenomenon of interest. The model learns an approximation $h^*(X)$ within a hypothesis space $\mathcal{H}$, where $\mathcal{H}$ represents the set of possible model behaviours $\{h_1, ..., h_n\}$. For this analysis, we assume ideal conditions where the training data accurately represents the population distribution, allowing us to focus on the core relationships between these functions.

The key distinction between intrinsically interpretable and post-hoc approaches lies in how they achieve human comprehensibility:

- In intrinsically interpretable models, $h^*(X)$ is constrained to be directly comprehensible, as indicated by the dashed boundary in the left diagram.
- In post-hoc models, an additional function $p^*(X)$ is learned from a space of possible explanations $\mathcal{P}$ to approximate $h^*(X)$'s behaviour in an interpretable way.

The learning paths (shown in red for intrinsic and blue for post-hoc) illustrate how each approach navigates the tension between model capability and human comprehensibility. While intrinsically interpretable models maintain a direct connection to human understanding, post-hoc methods introduce a secondary layer of interpretation that can potentially capture more complex relationships between the model and the underlying phenomenon.

B Theoretical Contexts and Epistemic Value in Machine Learning Interpretation

While Andrews [2023] demonstrates how theoretical considerations play essential roles throughout the ML pipeline in cases like AlphaFold – from data engineering to architecture design and model evaluation – not all AI models for science are built with this same level of theoretical rigour. CI expands this understanding by demonstrating that post-hoc interpretability methods can be epistemically justified across the full spectrum of theoretical contexts through structured mediated understanding. This broadens the scope of scientific ML beyond just theory-rich applications.

In theory-rich cases like AlphaFold, CI shows how post-hoc interpretability methods [Tan and Zhang, 2023] derive their epistemic validity from their role in a theoretically-grounded mediation process. They help validate whether models implement intended theoretical principles, enable scientists to verify if learned representations align with established theories, support discovery of potential new theoretical insights, and provide means to examine whether theoretical assumptions embedded in model design function as intended. The mediation process here acts as a bridge between explicit theoretical commitments and model behaviour, with interpretability methods serving as epistemically-justified tools for both validation and discovery precisely because they operate within a well-defined theoretical framework.

Even more significantly, CI demonstrates how post-hoc methods maintain epistemic validity in theory-poor contexts where ML models are applied with minimal theoretical consideration. Here, the mediation process takes on a different but equally valid epistemic role: helping uncover implicit theoretical assumptions embedded in data selection and preprocessing, providing ways to retrospectively examine model behaviour against domain knowledge, revealing potential biases or problematic patterns, and identifying opportunities for theoretical incorporation. Through mediated understanding, these methods create an epistemically rigorous path to reconstruct and validate theoretical frameworks that might have been overlooked in model development – effectively building theoretical bridges where none previously existed.

CI thus resolves a key epistemological challenge in machine learning interpretation: how to justify post-hoc interpretability methods across varying levels of theoretical grounding. The framework demonstrates that through properly structured mediation processes, these methods maintain epistemic validity whether they are validating existing theoretical frameworks or helping construct new ones. In both cases, post-hoc interpretability methods hold epistemic value because they enable systematic scientific scrutiny of how these models operate within their domains, regardless of how deliberately theory was incorporated into their development. This unified treatment of theory-rich and theory-poor contexts represents a significant advance in our understanding of how ML can contribute to scientific knowledge across diverse applications.

C Relevant Philosophical discussions on AI and Interpretability

C.1 Sullivan's (2022) Link Uncertainty

According to Sullivan, the relationship between explanation and understanding in complex models hinges critically on the concept of "link uncertainty" – the gap between a model's theoretical predictions and empirical reality. Sullivan argues that while models can be epistemically opaque (meaning their internal workings are not fully transparent), this opacity does not necessarily prevent them from providing genuine understanding, provided there is sufficient empirical evidence connecting the model to real-world phenomena. In other words, we do not necessarily need to fully understand how a model works internally; what matters more is understanding how the model connects to the real-world system it is studying.

Sullivan identifies three distinct types of explanatory questions we can ask about models: how the model itself works, how-possibly questions about potential mechanisms, and why/how-actually questions about real-world phenomena. How-possibly explanations demonstrate potential mechanisms or causes, showing how something could theoretically occur. However, these fall short of explaining how things actually work in reality. Using Schelling's segregation model as an example, Sullivan shows that while the model can demonstrate how segregation could emerge from individual preferences, it only provides genuine understanding if there is empirical evidence showing these mechanisms actually operate in real-world segregation patterns.

The decisive factor in moving from how-possibly to how-actually explanations is reducing link uncertainty through scientific evidence. This evidence must connect the model's theoretical insights to actual causal mechanisms in the target phenomenon. Importantly, Sullivan argues that understanding does not necessarily require complete knowledge of how a model works internally. Instead, what matters is the strength – whether it be the amount, kind or quality – of scientific and empirical evidence connecting the model's predictions or insights to real-world phenomena.

C.2 Andrews's (2023) Theory-ladenness of Machine Learning

The debate over Machine Learning's impact on science has generated what scholars call the "distinctness claim". The claim's core argument – articulated by several philosophers like Boge, Sreckovic et al., and Boon – is that ML, particularly deep learning, represents a fundamental departure from traditional scientific methods. They primarily base this on two key distinctions: (1) ML methods are supposedly "theory-free" or "theory-agnostic", operating without prior theoretical assumptions or conceptualisations of target phenomena, and (2) ML models prioritise prediction over explanation and understanding, making them epistemically opaque in novel ways. This perspective has gained significant traction not only in philosophical discourse but also among scientists and engineers who view ML as fundamentally different from traditional scientific approaches.

Extending Leonelli, Andrews fundamentally challenges this perspective with the theory-laden nature of scientific data and practice:

> Even the most simplistic of experimental designs reveals the nature and extent to which data, and scientific practice at large, are 'theory-laden.' The very act of investigation involves commitment to the existence and in-principle measurability of some phenomenon and, if we are making measurements and performing quantitative analyses thereon, commitment to its quantitative nature...
> Measurement cannot be total, and therefore there is always a commitment as to what to look at experimentally and what to exclude. There is always a commitment to the appropriate level of abstraction at which to study the phenomenon in play in terms of such things as instrument settings like degree of magnification or periodicity of sampling. The very design of our instruments of measure and their calibration includes various commitments to the nature of the worldly phenomena under investigation. [Andrews, 2023, pp. 6]

This understanding of data's theory-laden nature is now widely accepted in philosophy of science. However, as Leonelli notes, unfortunate relics of this view – viewing data as mere 'empirical input for modelling' – remain widespread. This persistent misconception underlies many arguments about ML's theory-free nature. The reality is that all scientific data, whether used in traditional methods or ML, necessarily involves theoretical assumptions and conceptual frameworks in its collection, preparation, and interpretation. This view challenges the technological determinism implicit in many discussions of ML in science – the belief that certain effects or limitations of ML are fixed, inevitable consequences of the technology itself. Rather than accepting current limitations as inherent features, Andrews argues we should recognise them as methodological challenges that can be addressed through improved practices and understanding.

Building on this theoretical foundation, Andrews demonstrates how the impossibility of "theory-free" learning is established by both philosophy of science and theoretical computer science's understanding of inductive generalisation. At its core, machine learning performs inductive inference - extrapolating from limited instances to general cases. Drawing on Norton's material theory of induction, Andrews notes that successful inductive inference never proceeds through universal, domain-generic formal rules, but rather requires the application of local rules warranted by empirical facts specific to each research domain. This philosophical insight finds independent confirmation in computer science through the No Free Lunch theorems, which mathematically demonstrate the impossibility of universal domain-generic inference rules. While these theorems were derived in specific formal settings, their implications for ML practice are profound: inductive inference fundamentally requires domain-specific inductive biases. This convergence of philosophical and mathematical results undermines claims about ML's theory-independence.

These insights reveal the "distinctness claim" as fundamentally misguided. Rather than representing a revolutionary break from traditional scientific methods, ML should be understood as a new set of tools whose proper application still requires theoretical understanding and methodological rigour. This perspective suggests a more nuanced approach to ML in science: one that acknowledges how theoretical considerations may enter differently in ML workflows, while recognising their essential role in ensuring sound scientific practice. Such an understanding is crucial for developing appropriate methodological standards for ML in science, rather than accepting current limitations as inevitable features of the technology.

C.3 Beisbart and Räz's (2022) Factivity Dilemma

The factivity dilemma in understanding Deep Neural Networks (DNNs) centers on a fundamental tension between accuracy and comprehensibility. The principle of factivity demands that explanations and understanding be grounded in facts, yet modern DNNs have become so complex that we can only comprehend them through simplifications and idealizations. As Rudin (2019) pointedly argues, a perfectly accurate explanation would simply duplicate the original model's complexity, defeating the purpose of explanation. This creates what appears to be an insurmountable challenge: explanations must either sacrifice accuracy for comprehensibility or maintain accuracy at the cost of being unusable.

This tension has deep roots in the philosophy of science, particularly in debates about the relationship between explanation and understanding. Traditional accounts of scientific explanation, such as the Deductive-Nomological model, typically require factivity - the premises in an explanation must be true. However, the requirements for scientific understanding are more nuanced. Non-factivists like Elgin argue that simplified models can provide legitimate understanding despite imperfect accuracy, while factivists such as Lawler maintain that simplifications are merely instruments toward understanding rather than constituting understanding itself. These opposing views reflect a broader debate about whether understanding necessarily requires truth or can be achieved through useful approximations.

A potential resolution emerges when we distinguish between mechanistic interpretability and scientific understanding in the context of DNNs. While mechanistic interpretability aims for factual explanations of model behaviour, scientific understanding of phenomena through post-hoc interpretative models may not require the same level of factivity. This distinction suggests that while complete, accurate explanations remain an important goal, we can develop meaningful understanding through carefully constructed simplified models. The key lies in maintaining awareness of these models' limitations while leveraging their insights - acknowledging them as useful approximations rather than complete representations of reality. This approach offers a practical way forward, recognizing both the current constraints in explaining DNNs and the necessity of working with these systems, even with imperfect understanding.

C.4 Freiesleben et al.'s (2024) Holistic Representationality

Freiesleben et al. address a fundamental challenge in modern scientific research: how to derive meaningful scientific insights from machine learning models that, unlike traditional scientific models, lack direct interpretability of their components. Traditional scientific models followed what the authors call "elementwise representationality" (ER), where each model component – whether parameters, variables, or relationships – directly represented something meaningful about the phenomenon being studied. For instance, in a simple physics model, mass and velocity parameters directly correspond to physical properties. However, modern ML models, particularly neural networks, do not offer this kind of straightforward interpretation - their individual components (like network weights) do not map clearly to real-world phenomena (e.g., see Freiesleben [2023]).

Rather than viewing this as a limitation, the authors propose a framework based on "holistic representationality" (HR). Instead of trying to interpret individual components, they suggest analysing the model's behaviour as a whole through what they call "property descriptors" (e.g., cPDP, cFI, SAGE, and PRIM for global property, and ICE, cSV, ICI and Counterfactuals for local property, see pp.21-25 for further details). This approach aligns with recent findings from Bilodeau et al. [2024], who demonstrate that generic feature attribution methods can be unreliable for inferring model behaviour, but task-specific approaches can dramatically improve interpretability. While Freiesleben et al. provide a theoretical framework, Bilodeau et al. offer practical evidence of its importance, showing how domain-specific interpretability methods (e.g., perturbation) can be more reliable than general-purpose approaches like SHAP or Integrated Gradients.

The authors provide a systematic four-step framework for this approach (pp. 14-20): first, formalising the scientific question as a statistical query, which involves translating research questions into precise mathematical formulations; second, identifying how to answer it using the whole model through property descriptors that are continuous functions mapping from model space to answer space; third, estimating the answer using the trained model, which requires careful consideration of data distribution and model behavior; and fourth, quantifying the uncertainty in the results through both model error (difference between optimal and trained model predictions) and estimation error (uncertainty in the property descriptor estimates themselves). This framework is particularly notable for its rigorous treatment of uncertainty quantification, which is often overlooked in traditional interpretable ML approaches.

The paper demonstrates the practical applicability of this framework by showing how existing interpretable ML methods can serve as property descriptors. Using a concrete example of analysing student academic performance, they illustrate how these methods can provide scientifically meaningful insights while maintaining rigorous standards of inference. The authors emphasise that while this approach differs from traditional scientific modelling, it does not sacrifice scientific rigour – it simply provides a different path to extracting knowledge from our models, one that's better suited to the capabilities and limitations of modern machine learning systems. This conclusion resonates with Bilodeau et al.'s findings that success in model interpretation often depends on carefully defining concrete end-tasks and developing targeted evaluation methods rather than relying on general-purpose interpretation tools.

C.5 Lazar's (2024) Democratic Duties of Explanation

Lazar's central contribution to AI explainability discourse stems from his recognition that computational systems, especially AI, are increasingly being used to "govern" us – that is, to settle, implement, and enforce the norms that determine how institutions function. When computational systems are deployed by government agencies in administrative functions or by private companies to police online behaviour and determine our information access, they are effectively governing us. For such governing power to be legitimate, Lazar argues, it must be accountable to democratic oversight through public explanation to the community as a whole. Unlike approaches focused on individual rights or technical transparency, Lazar emphasises that explainability is fundamentally a democratic duty – it is not about individual decision subjects understanding their particular outcomes, but about enabling the collective community to determine whether these computational governance systems are being used legitimately and with proper authority. Lazar argues that this collective explainability requirement has specific implications for computational governance systems: they must reveal not just their decision rules, but also demonstrate the appropriateness of their training data as evidence, the robustness of their decision-making processes, and their ability to make the right decisions for the right reasons.

C.6 Vredenburgh's (2022) Informed Self-advocacy

Vredenburgh's central contribution addresses the fundamental tension between algorithmic opacity and individual rights. Rather than demanding complete technical transparency of complex AI models, she argues for a claim right to explanations that can be provided post-hoc, grounded in what she calls "informed self-advocacy" – a cluster of abilities that allows individuals to represent their interests and values to decision-makers and further those interests within institutions. This right becomes particularly crucial in institutions where algorithmic decisions significantly impact individuals' lives.

Vredenburgh argues that post-hoc explanations must take two specific forms: rule-based normative explanations (explaining why a decision was appropriate) and rule-based causal explanations (explaining how inputs relate to outputs). She advocates for "functional transparency" – high-level explanations of how inputs relate to outputs – rather than structural or run transparency of the underlying model (pp. 13). While acknowledging that simplified explanations of complex algorithms may be somewhat inaccurate, she argues they can still be sufficient for informed self-advocacy if properly calibrated to stakes: when decisions distribute harms or entitlements (versus benefits), there are stronger requirements for clear explanations and human expert support. This pragmatic framework shows how post-hoc explanations, even if they do not fully capture the complexity of AI systems, can satisfy legitimate needs for accountability while remaining feasibly implementable, as evidenced by existing legal requirements for explanation across various domains.

C.7 Durán's (2023) Computational Realibilism

The central motivation of justification in DNNs is primarily driven by their inherent methodological and epistemic opacity [Humphreys, 2009]. This opacity manifests in two distinct yet interrelated ways [Durán, 2023]. First, the algorithmic complexity of DNN systems – encompassing myriad functions, variables, decisions, and data – renders it impossible for any individual or group to fully comprehend which elements are pivotal in generating a specific output. Second, this complexity imposes cognitive limitations on human agents, hindering our ability to derive meaningful interpretations of the algorithm and its results. Both aspects of opacity potentially undermine the justificatory basis for ascribing scientific value to DNN outputs, either due to the "black-box" nature of the system or the cognitive constraints of human interpreters.

Duran's computational reliabilism (CR) addresses this epistemic challenge by proposing a framework for justifying belief in DNN outputs if and when they are produced by reliable belief-forming methods [Durán and Formanek, 2018, Durán, 2023, Javed et al., 2023]. CR delineates three categories of reliability indicators: (i) Technical Robustness of Algorithms, encompassing the design, implementation, and maintenance factors that contribute to a DNN system's robustness; (ii) Computer-based Scientific Practice, which involves the algorithmic implementation of scientific theories and principles, or expert assessment within established scientific knowledge; and (iii) Social Construction of Reliability, referring to the socially mediated processes that confer acceptance of DNN and its outputs across diverse communities. At its core, CR adopts a frequentist approach, positing that beliefs formed by demonstrably reliable algorithms warrant greater justification than those produced by unreliable ones.