Philosophical Analyses of ML Modelling in Science

A thesis submitted in partial fulfillment of the requirements
for the
BS-MS dual degree
Indian Institute of Science Education and Research, Pune
March, 2025

by

Saransh Agrawal

Supervisor: ***Prof. Varun Bhatta***
Indian Institute of Science Education and Research, Bhopal (IISER-B)

Expert: ***Prof. G. Nagarjuna***
Indian Institute of Science Education and Research, Pune (IISER-P)

भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान पुणे
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

IISER PUNE

Certificate

This is to certify that this thesis entitled *"Philosophical Analyses of ML Modelling in Science"* towards partial fulfillment of the BS-MS dual degree program at the Indian Institute of Science Education and Research, Pune, represents work carried out by Saransh Agrawal at IISER, Bhopal under the supervision of Prof. Varun Bhatta during the academic year 2024-2025.

Prof. Varun Bhatta
(Supervisor)

Prof. G. Nagarjuna
(Expert)

Committee:
Prof. Varun Bhatta
Prof. G. Nagarjuna

भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान पुणे
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

IISER PUNE

Declaration

I hereby declare that the matter embodied in the thesis entitled ***"Philosophical Analyses of ML Modelling in Science"*** is the result of work carried out by me at IISER, Bhopal under the supervision of Prof. Varun Bhatta and has not been submitted elsewhere for any other degree. Wherever others have contributed, every effort has been made to indicate this clearly with due reference to the literature and acknowledgment of collaborative research and discussions.

Saransh Agrawal
(20201236)

This thesis is jointly dedicated to:

My parents,
for all their love and support

&

K.P. Mohanan,
for empowering me to move beyond merely filling pails of knowledge and giving me the tools for lighting proverbial fires.

ACKNOWLEDGMENTS

First and foremost, I want to thank my supervisor, Prof. Varun Bhatta, for his exceptional academic guidance and personal support throughout this thesis. Without Varun, I would not have been able to transition from pursuing research in science to scientific epistemology while remaining in India. I am deeply grateful for his constant insistence on rigor, clarity, and substance. Varun has painstakingly worked to correct numerous mistakes and issues throughout this thesis (there have been many) and any mistakes that remain are entirely due to me.

I also thank Prof. G. Nagarjuna, for agreeing to serve as my expert and for the valuable discussions related to the thesis. I am especially grateful to him for mentoring me in a semester long reading project that gave me the confidence to pursue my MS in scientific epistemology. I extend my thanks to our Dean of Academics, Prof. Girish Ratnaparkhi, for his support to my thesis, inspite of the many challenges and problems that I have posed for him. I am also grateful to Prof. Joy Monteiro for his personal guidance and academic discussions, both of which have contributed to improving this thesis.

A special note of acknowledgement is due to Sundar Sarukkai, whose encouragement allowed me to seriously consider pursuing philosophy of science in India. His life's work has paved the way for younger scholars like myself to explore the fascinating questions in this field. I am deeply grateful for his extensive list of published works, which has shaped my understanding of philosophy of science. Moreover, I am also indebted to his discussions regarding the thesis, and the personal support and guidance that he has generously shared with me.

I have dedicated this thesis to K.P. Mohanan. Mo has been my mentor in numerous personal challenges and academic pursuits, and has been instrumental in my development as a person. I am forever indebted to him and hope to repay the knowledge, kindness, and wisdom that he has shared with me by paying it forward. I hope to make a meaningful impact on other people through my research and teaching, just as Mo has

done for me.

Moreover, without Mo, I would not have found the incredible community at ThinQ, including Tara, Vignesh, and Rahul, nor would I have had the opportunity to host my dear CICL sessions. My intellectual discipline and passion for research are shaped by Mo, as is the foundation of my public academic life.

Lastly, I would like to express my gratitude to my family and friends for their unwavering love and support. Vivek, Vihang, Krish, Aditya, Sayan, Aaditya, Naman, Gaurav, Sraavya, Harshitha, PK, Avneet, Vinayak, Vatsal, Sumanth, Akilan, Epsita, Manav, Sabareesh, and my colleagues in Bhopal—Shobha, Rupesh, GD, Glincy, and many others. Many of them have also engaged with me in extended academic discussions that have enriched this thesis.

My family in Nagpur has showered me with immense love and support, which has sustained my work more than anything else.

ABSTRACT

The widespread adoption of Machine Learning (ML) and Artificial Intelligence (AI) in scientific practice has raised novel philosophical questions concerning the epistemic status of these technologies. In this thesis, I will examine the property of *epistemic opacity* (also referred to as the "black-box" problem), which poses significant challenges for the users of these technologies. I provide an argumentative literature review of epistemic opacity in AI-ML models, and analyze how the black-box nature of these technologies undermines the epistemic goals for which these models are deployed. Unlike the theoretically grounded models of conventional scientific practice, ML models make inferences by identifying statistical correlations in the data itself. Although a ML model might make accurate predictions, even the scientists who have constructed the model might lack access to the "inner workings" of the ML model. This is because the scientists lack a direct theoretical interpretation of the epistemic components of a ML model—for instance, the significance of the weights assigned to a set of parameters constituting a neural network. This raises questions about the epistemic justification for using ML techniques in scientific practice. Moreover, this has also led to widespread debate concerning the trade-offs between predictive capabilities, explanatory value, theoretical understanding, and other epistemic desiderata for working scientists. I aim to contribute to this debate by highlighting the plurality of meanings attributed to fundamental scientific concepts like prediction and discovery and argue for the utility of distinguishing between different conceptual notions that are associated with these terms. Furthermore, I also argue that *discovery* and *prediction* claims in ML modelling rely on different modes of justification compared to conventional scientific practice and how these different modes of justification can shape the meaning taken up by the concepts of *discovery* and *prediction* in the context of ML modelling in science.

Contents

CHAPTER 1

INTRODUCTION

Technologies constitute an important factor that shape the conduct of science. In the past few decades, computer simulations have been integrated into various domains in science. More recently, AI (Artificial Intelligence) technologies like ML (Machine Learning) models are also being widely used by scientists across different fields. This is evident by a *Nature* survey of 1600 researchers wherein the majority of scientists believe that AI technologies will soon become central to the practice of research (Van Noorden and Perkel 2023). These new technologies provide various affordances to scientists by providing faster ways to process data, automating repeated tasks, helping in writing code, and so on. However, these technologies also have certain properties that makes it hard to situate them in the conventional scientific method.

In this thesis, I seek to understand how the adoption of new AI technologies impacts the conduct of science. My research is situated in philosophy of science, and not philosophy of mind or philosophy of AI. My primary focus will be on identifying the impact that the adoption of these novel technologies might have on science, rather than the very nature of these technologies. Towards this end, I will first do a literature review of the epistemology of computer simulations in science. Subsequently, I will move to a more recent phenomena of the widespread adoption of AI technologies by scientists. AI is a very broad term, and in this thesis, I will limit my focus to a subset of AI that is referred to as connectionist AI models (more specifically, ML models).

Suppose that an expert modelling a scientific phenomena is attempting to understand the inner workings of a successful ML model that she has built, for example, using a neural network architecture. In the course of this process, it is true that the expert can peek inside the model and gain access to different model parameters and the respective weights that the learned model has assigned to them. However, the expert will face difficulty in finding any direct semantic interpretation of the modelled phenomena solely using the weights and parameters of the ML model. It is as if these models are

using a different language for describing the data (and by proxy, the phenomena) that is not humanly understandable.[1] This property is that of *epistemic opacity*—colloquially referred to as the "black-box" nature of certain computer programs—a property that is common to both computer simulations and AI technologies. I will argue that the epistemic opacity of these technologies has important implications for any epistemic activity in which these technologies are employed.

The philosophical inquiry proposed in this thesis may encounter a skeptical challenge: If scientists claim that ML models are "working out" for them, what is the need for any philosiphizing? Can't we say that the "proof is in the pudding" as ML models are able to make accurate predictions based on domain specific data sets? In response, I will try to demonstrate that even if these ML models are making accurate predictions, this might still not satisfy the epistemic demands of science. Even if scientists are able to successfully train a ML model to make accurate predictions on the basis of a data set, the scientists may still want to determine whether a model is employing a heuristic, whether the correlation patterns are a mere artifact of the data set, or whether a model has captured aspects of the deeper causal structures that could have manifested the data. The property of epistemic opacity complicates this problem because scientists do not have access to the inner elements of a ML model which can help them determine why the ML model is able to make successful predictions.

Simply put, even if an AI model is able to make accurate predictions, a scientist might still not know whether:

1) The AI *truly understands* something. Here, true understanding would mean that the AI is able to *learn* the relevant categories or concepts in the data.[2]

2) AI gaining a *superficial understanding* based on statistical associations without developing a true understanding of the data.

In the context of ML modelling, 1) would correspond to a ML model developing *emergent abilites* that can move beyond the constraints of a training data set. On the other hand, 2) would correspond to a practice of making novel predictions by using a method of *approximate retrieval* from the information stored in a training data set. Put differently: are the predictive capabilities of a ML model due to its *causal reasoning abilities*, or due to the *contextual heuristics* of a particular data set. If it is the latter, what is the epistemic status of these representations and statistical correlations inside an ML model? Are they instances of knowledge, are they heuristics, or can they be qualified as proper philosophical categories or scientific concepts?

1. My understanding of philosophy of science, and the idea of language and mathematics as a description of reality is greatly influenced by the works of Sundar Sarukkai, especially his book titled 'what is science?' (2012).

2. Thanks to K.P. Mohanan for extensive discussions on this topic.

Although a thorough treatment of these sets of questions is outside the scope of this thesis, I will focus my attention on analyzing the meaning of fundamental concepts like explanation, prediction, and discovery in the philosophy of ML modelling in science. I will argue that conceptual clarification of these words will help us address the challenges associated with the rising adoption of AI in science. For instance, if these ML models were to continue outperforming intelligible and theoretically grounded scientific models in terms of predictive capabilities, will this mean that conventional scientific models will be rendered obsolete?

The question brings out foundational assumptions about the epistemic goals for which these models are being developed. Here, I will introduce the concept of *epistemic desiderata* as opposed to epistemic goals. In my usage, epistemic desiderata are virtues that contribute to epistemic goodness like predictive capabilities, coherence, explanatory power, and so on. Aiming for epistemic desiderata help us achieve the broader epistemic *goals* such as those of gaining knowledge and understanding. The assumption seems to be that predictive capabilities are a sufficient epistemic desiderata in scientific modelling even if it comes at the cost of coherence, explanatory value and theoretical understanding. But is it even possible to ever make a predictive model that has no explanatory value? Can there exist predictions that do not explain anything? And if so, is it desirable to construct theories and models that can make predictions but do not provide any explanatory value?

Some philosophers say yes. Srećković, Berber, and Filipović (2022) posit that the rising adoption of ML technologies has led to a disruption in conventional scientific method because these models have started making "predictions without explanation". This is because ML models have the ability to make inferences solely using data in a "theory-agnostic" or even a "theory-free" manner. Srećković et al.'s argument is one side of a wider debate concerning the concepts of prediction and explanation in the philosophy of ML modelling in science. However in my review of the debate, I find that the concepts of *prediction* and *explanation* are being used in a plurality of ways by scholars. I think that one reason for this is due to the historical context of these concepts being defined on the basis of theoretically grounded models and conventional scientific method. Some philosophers posit that this novel data-driven method of making inferences using ML (and big data science in general) is fundamentally different from conventional scientific methods (Leonelli 2020).

This motivates the need to review a triad of the fundamental concepts of explanation, prediction, and discovery, and their usage by scholars in ML modelling in science. I will argue that unlike explanation, philosophers of ML modelling in science have not paid sufficient attention to the concepts of prediction and discovery.

In my review of this debate, I came upon another problem causing widespread confusion among scholars. This issue was about the existence and nature of the trade-

offs between various epistemic desiderata in ML modelling. For instance, is there necessarily an epistemic trade-off between predictive capabilities and explanatory value? Philosophers are divided on the issue, and I think one reason behind the division is the plurality of meanings taken up by the words in play. Moreover, as I shall demonstrate in the final chapter, scientists publishing review articles on the nature of the trade-off cannot come to any conclusion either. Confusion is rampant, and I will argue that one reason for this is the ambiguous usage of prediction as a concept in the work of scholars; philosophers and scientists alike.

It is possible that words like explanation, prediction, and discovery take on different meanings in the context of ML modelling. And if ML and other AI technologies can continue to make headway in their epistemic capabilities, we just might have to change the way we use these words in science proper. In this vein, I will put forward an argument to support the thesis that these fundamental concepts like prediction indeed have different meanings in the context of ML modelling. I will examine the different modes of justification behind the *outputs* of an ML model that allow scientists to accept these outputs and qualify them as *predictions*. The output of a model can only be called a prediction based on certain properties of the model; here, the model is being conceptualised as an epistemic process that generates knowledge claims. The properties of the model that lets us trust the results of the model are the reliability conditions of this epistemic process of generating knowledge claims.

We can only justify the outputs of certain ML models by comparing them with empirical data because of a lack of theoretical understanding into the inner workings of the ML model. This is unlike conventional science, where predictions are grounded in theory. Any new discoveries or prediction can gain additional support if it is found to cohere with the larger theory of a domain. However, such a mode of justification is not available to ML models.

In chapter 2, I will give a review of the epistemology of computer simulations. This is followed by a deeper analysis of the concept of epistemic opacity in chapter 3 which is a common thread running across this thesis. I first introduce the concept of epistemic opacity in the context of computer simulations, and then proceed to examine its usage in the context of AI and ML.

I give a very brief review of AI and ML in chapter 4. In chapter 5, I review the concepts of explanation, prediction, and discovery, by comparing how these concepts are understood in general philosophy of science versus ML modelling in science.

I have made a conscious choice to withold my discussion on method until I start framing my central argument in chapter 6. This is done to ensure that methodological discussions are only presented where they are relevant, and serve to justify my chosen approach. In chapter 6, I try to demonstrate the ambiguous usage of the concept of prediction and discovery by scholars of ML modelling in science and review some

arguments about the changing nature of the concept of prediction with the advent of AI technologies. Finally, in chapter 7, I argue that predictions made by certain ML models are different from the predictions of conventional science because of the limited modes of justification that are afforded to experts employing ML models. I end the chapter by giving a comprehensive meta-review of various review articles from different scientific domains to highlight the confusion about the trade-offs between various epistemic desiderata in ML modelling. I will further argue that this confusion is (at least partly) a result of the ambiguous usage of concepts like explanation and prediction.

CHAPTER 2

EPISTEMOLOGY OF COMPUTER SIMULATIONS IN SCIENCE

In a short chapter titled 'The End of Insight,' Steven Strogatz states that:

> ...there are simple computer programs...whose dynamics can be so inscrutable that there's no way to predict how they'll behave. The best you can do is simulate them on the computer, sit back, and watch how they unfold. Observation replaces insight. Mathematics becomes a spectator sport (Strogatz 2007, 131)

Strogatz goes on to express his concern about the shift in contemporary mathematical and scientific practice due to its increasing reliance on computers. Similar claims about shifts in scientific practice by the adoption of novel technologies are common. More recently, scholars have posited how the introduction of big data algorithms can undermine the need for domain-specific expertise into causal mechanisms that are behind the correlational results in a large data set (Andrews 2024a). However, even if the claims might be warranted, we should be skeptical about proclamations anticipating paradigm-shifting revolutions with the introduction of a new technology in science.

Pen and paper methods, limited as they are in their scope of application, can nevertheless be exactly true to theory in terms of potentially providing exact analytical solutions. However, a computer program must rely on approximation methods for even the simplest of arithmetic problems like $0.1 + 0.2$.[1] If a computer is incapable of representing a perfectly exact answer for such simple arithmetic operations, how can we trust its results when it is running a model that is calculating millions of differential equations simultaneously?

1. This occurs because all decimal numbers cannot be exactly represented in binary form (Fahim 2024). A computer cannot handle an infinite series representation of a number and has to necessarily truncate the value to a particular degree of accuracy.

Despite these skeptical concerns, computer program (which includes computer simulation models) yield results that scientists deem to be reliable knowledge is vastly complex situations. The Wolfram project in physics—which suggests that computer science is the correct foundational methodology in constructing the theory of everything in physics—is another example of the kind of trust that some scientists place in computer programs.

Computer simulation models are a subset of the computer programs that Strogatz (2007) is referring to. In the subsequent sections of this chapter, I will review a set of scholarly works that situate computer simulation in the pre-existing philosophical frameworks around the concepts of modelling, experiments, and representation. I will explore questions like – how do we trust the results of a computer simulation model? What is the relationship between computer simulations and experiments? What questions and problems do computer simulations models raise for philosophy of science? This will allow me to demonstrate the significance of the concept of epistemic opacity that I will highlight in the final sections of this chapter.

2.1 Defining computer simulations

The list of sciences that use computer simulation ranges from physics and engineering to medicine and sociology (Winsberg 2022). But what exactly are computer simulations in science?

In contrast to use of the word simulation in ordinary language, the result of a computer simulation need not be an image or an animation on a computer screen that is comprehensible to the user. The purpose of a computer simulation algorithm is to process input data and convert it into more useful output data. It is only this output data that is (sometimes, but not always) represented using visual tools for human comprehension.

A more technical definition of computer simulation conceptualises it as a systematic process of studying real-world systems (these systems can also be fictional, like those represented by a mathematical model) by modelling them in a form that can be run on a computer in the form of a computer program. These mathematical models are made more amenable to calculation by various techniques of approximation, for example, by using numerical methods for solving differential equations (Winsberg 2022).

Building on this definition, Winsberg states that computer simulations model real-world systems by idealizing them using several variables of interest. A collection of variables like temperature, pressure, and so on, will combine to describe a particular state S_1 of the system at a time T_1. The state S_1 at time T_1 evolves to state S_2 at time T_2, and so on. The goal of studying the computer simulation model is to understand the

rules of evolution that govern this dynamic behavior.

A narrower definition of computer simulation refers to the set of algorithms that are run on a particular computer using a particular compiler. However, a broader definition of a computer simulation includes the entire process of constructing an algorithm, the additional affordances that the computer simulation model provides, and its implementation by scientists in their work (Winsberg 2022).

For Beisbart (2023), a computer simulation model is used to trace the state of a system as it evolves in time, for instance, a dynamic computer simulation of a tsunami as it approaches a coastal city. The emphasis on a system's "time evolution" highlights how the modelling practice has its roots in the physical sciences (Beisbart 2023). However, the use of computer simulation models has now expanded to various other fields, and it will be fruitful to consider alternative descriptions that will allow us to expand the range of computer models that we can study as a singular category. One way to do this is to reconsider the inclusion of the time evolution of a system as a central criterion to categorize a set of algorithms as a computer simulation model. However, even this basic change is challenged by philosophers, some of whom seem to explicitly contest this:

> ...not every use of the computer qualifies as computer simulation; for instance, the classification of images using neural networks does not count as computer simulation because no time evolution is traced. (Beisbart 2023)

Computer simulations are categorized into a few different types (E. Winsberg 2022). Some of these are:

1) *Equation-based* simulations, where certain universal laws of evolution govern the dynamics of individual particles. These simulations are prevalent in physical sciences, for instance, a model of a fluid flowing through a pipe.
2) *Agent-based* simulations, where individual, autonomous entities called agents are modeled in the computer simulation model. The dynamics of these agents are governed by their immediate local environment rather than any general universal rules. These models are often used in social sciences, for example, to model the behavior of a population.
3) *Multiscale* models combine elements from different scales of the target system. For example, consider the scientific problem of measuring large-scale ocean circulation patterns. Given the enormity of this task, one can choose to reduce the computational cost of the effect of surface wind on ocean mixing. An eddy viscosity parameter can be added to represent the extra turbulent mixing resulting from the wind. In this way, small-scale processes are simplified into a few manageable parameters and exported to an algorithm that describes the larger-scale dynamics.

Parameterization, a concept that will be explored later, is a multiscale modelling technique widely used in climate models.

However, representing complex lower-scale models with a set of parameters raises critical questions about representation, modelling, and the justification for this substitution. Are we merely replacing intricate processes with a simplified mathematical description? How exactly does this happen? In what contexts can we successfully adopt this strategy? Sometimes, these models are constructed in a manner where different parameters are derived from different, incompatible theories. This situation prompts questions about how these incompatible theories, which cannot directly speak to each other, can coalesce to produce results that seem to align with empirical data (Winsberg 2022).

2.2 Epistemology of computer simulations

Winsberg (2022) discusses Schelling's 1971 model of "segregation," which demonstrates how a slight preference for residing next to a person of one's own race can lead to large-scale patterns of segregation. These large-scale patterns were similar to the settlement patterns of black people and white people in American cities. However, it is unclear as to what epistemic desiderata is provided by this model. Is this a possible-explanation of the large-scale pattern that we see in the real world? Is it predicting future outcomes, or is it reproducing observed phenomena?

The epistemology of computer simulations aims to establish a framework for understanding the foundations of computer simulation and its relevant epistemic goals, such as prediction, explanation, understanding, and so on (Winsberg 2022). Given the increasing reliance of modern science and technology on computer simulations for designing airplanes, forecasting storms, and shaping climate policies, it is crucial to discern whether computer simulation models differ from conventional modelling practices and to determine the epistemic warrants for our belief in the results of a computer simulation.

Winsberg (2001, as cited in Winsberg 2022), gives an account of three conditions that must be met for an adequate epistemology of computer simulation. These conditions refer to the kinds of inferences that are generally drawn by a computer simulation. These are:

1) *Downward*: Model construction in computer simulation is guided *top-down* from theory rather than *bottom-up* from observation.[2] Theory usually acts as the starting point for formulating computer simulation models.

2. This is a common theme that will be explored in the thesis.

2) *Motley*: Modelling in computer simulation doesn't rely on theory alone but also on computer libraries, mathematical and computational techniques for approximating equations, and so on.
3) *Autonomous*: Often, the results of a computer simulation cannot be directly compared with observations due to the difficulty in obtaining the same kind of knowledge from conventional experiments.

However, Parker (2014) notes that this set of conditions might be biased towards simulations in physical sciences. She further notes that a general epistemology of computer simulation should also incorporate the condition that inferences about future, as yet unexplored systems, may require more demanding justificatory strategies than inferences about how already observed features of a target system could be (re)produced.

A central question regarding the results of a computer simulation is whether it captures some part of reality like an experiment would. Another related question is whether a computer simulation can ever perform a measurement (Winsberg 2022). As experiments are prone to error from various sources, like improper calibration of instruments, impure samples, and so on, computer simulations might also be prone to new sources of errors like corrupted code, bugs, and imported packages comprising millions of lines of code that cannot completely be verified by a single individual.

2.3 Novel experiments, or mere tools for calculation?

Experiments serve as the crucial link that bridges theory and observation in science. However, it is difficult to situate computer simulation on the methodological map that outlines the relationship between theory and experimental outcomes (Barberousse 2018).[3]

Suppose a physicist runs a computer model to simulate the behavior of Earth if the ozone layer were to be depleted. By doing so, has she performed an experiment? The computer simulation lacks the direct physical intervention that is characteristic of common intuition regarding experimentation in science. However, the simulation model does share certain similarities with experiments due to the shared epistemic desiderata of testing hypotheses and discovering new phenomena. In its simplest form, the question is whether computer simulation are instances of experiments?

Hacking's described experiments as having a "life of their own", and substantiation of this claim vindicated of the status of experiments in philosophy of science (Winsberg 2022). Winsberg takes this further and attempts to draw parallels between the status of experiments and computer simulations. Computer simulations could happen to share

3. In this chapter titled 'Philosophy of Physics', Barberousse highlights the challenges that has been posed by the introduction of computer simulation in physics.

some essential properties of experiments, or on a more stronger condition, computer simulations could very well be instances of experiments. If this were true, then we could draw from the literature on experiments to justify our trust in the results of computer simulations.

Parker (2014) counters this claim by stating that computer simulations lack rigor and that their results cannot be directly compared with those of experiments. This is because computer simulation modelling mostly emphasizes comparing model results with empirical results, and very little effort is spent in investigating whether the model can provide evidence for testing scientific hypotheses (Parker 2014).

Winsberg (2022) attempts to categorize the diversity of views regarding computer simulation and their status as experiments in science by proposing two theses. The first is the *identity thesis*, which states that computer simulations are literally instances of experiments.[4] The second is the *epistemological dependence thesis*, which posits that the epistemological warrant for our belief in the results of a computer simulation relies on the identity thesis being true. However, the epistemological dependence thesis can be constructed with varying strengths, and alternate accounts for justification in the epistemology of computer simulation need not rely on computer simulation being instances of experiments (Winsberg 2022).

Critics of the view that computer simulations are instances of experiments highlight the dependence of a computer simulation on its underlying theory. They claim that a computer simulation is entirely situated within its foundational theory, and can only make logical inferences within the bounds of that theory. This is in contrast to an experiment which can surprise scientists by discovering novel results that resist explanation from accepted theory. In this vein, critical scholars posit that computer simulation models serve as mere tools. These tools are used to articulate new arguments that were not immediately available to our cognition by *extending* theory (E. Winsberg 2022). For instance, utilizing the available data-set of temperature and pressure readings to simulate a climate model and making predictions in the form of weather forecasts.

However, it is crucial to emphasize that often a scientist will have to actually run the model to know its results. The results of simulation models cannot preemptively be inferred by a set of logical steps. Similar to experiments, computer simulations are also capable of surprising us, which has led some philosophers to conceptualize them as experiments in their own right (Winsberg 2022).

Another distinction that is often drawn between experiments and computer simulations underscores the difference between the process of intervening on a model of the target system (for instance, a climate model simulating the Earth's atmosphere) versus intervening on the actual target system itself (a part of the Earth's physical atmosphere).

4. If computer simulation were experiments, then they would logically inherit all of the epistemological virtues enjoyed by experiments.

Some philosophers argue that computer simulation can only intervene on the model of the target system and not on the actual target system. According to this view, if a computer simulation lacks the ability to intervene on the actual target system, then it fails to meet the necessary criteria for being categorised an experiment (Winsberg 2022).

However, critics of this view respond by noting that this distinction may not be as straightforward in the case scientific experimentations themselves (Winsberg 2022). When we are conducting a concrete experiment, the process constitutes intervening on a model system that serves as a representation of the actual system of interest. There is a difference between the data that is obtained from an experimental setup, as opposed to the phenomena that is inferred using the experimental data (Daston 2008).

Winsberg (2022) bridges these contrary viewpoints by suggesting that the difference between computer simulations being treated as part of theory or experiment is primarily a simple shift in emphasis, rather than a fundamental issue that significantly impacts broader questions in philosophy of science. Similarly, Parker (2014) argues that whether we classify computer simulation as experiments or not may not have substantial implications for philosophy of science. This is because the process of labeling them as sharing some properties of experiments does not automatically imply that they are epistemically identical to conventional experiments. Moreover, identifying computer simulation as experiments does not imply that they can completely replace conventional experiments in any way. This is due to the obvious fact that certain experiments, such as determining the hardness of a new material using the Mohs scale, necessarily require physical experimentation.

Furthermore, Barberousse (2018) posits that rather than locating computer simulation closer to theory or experiment on the methodological map, what is more important is to understand the validity of computer simulation, and our warrants for trusting its results.

2.4 Verification and validation

Practitioners employing computer simulations, such as scientists and engineers, often make a distinction between what they call verification and validation (Winsberg 2022). *Verification* refers to the process of bringing the results of a computer simulation model within a close range to the foundational equations of the model. *Validation* refers to the process of testing the validity of using a particular model to represent the target system of interest, for instance, by comparing the results of a computer simulation with experimental data from the target system. Verification involves debugging and implementing mathematical techniques, while validation relies on direct comparison with real-world data.

Frigg and Reiss (2009, as cited in Winsberg 2022) utilize the verification and validation distinction to categorize the supposedly unique aspects of computer simulation. They argue against claims of philosophical novelty of computer simulations by trying to situate all novel aspects of simulations as being either mathematical (verification), or being subsumed under various accounts of modelling practices (validation).[5]

Despite the popularity[6] of this usage, Winsberg (2022) argues that these practices of verficiation and validation cannot be cleanly separated. He posits that verification is not purely mathematical by highlighting instances in computer simulation modelling where scientists use a set of equations that do no have an exact analytical solution. He further argues that scientists often select equations that have the advantage of being more tractable (verification) rather than those that would be ideal if we were to derive them from theory (validation). This clearly shows how both verification and validation are linked in practice.[7]

Barberousse (2018) argues against the utility of the verification and validation distinction by highlight how it is very difficult for an individual scientist to verify the code behind a computer simulation model.[8] Furthermore, Barberousse states that it is not particuarly productive and illuminating to focus on whether computer simulation models have false assumptions at their base.[9] Instead, the focus should be on testing hypotheses in specific contexts by comparing model results with empirical data.

Although some practicing scientists describe computer simulation as instances of experiments,[10] only the validation of model results with empirical data can significantly bolster the credibility of a computer simulation (Nersessian 2022). However, simulations frequently model scenarios where conducting actual experiments is impossible, such as assessing the impact of detonating a nuclear bomb over New Delhi.

2.5 Philosophical novelty of computer simulations in science

Suppose we happen to find novel features in the epistemology of computer simulation that are not part and parcel of conventional modelling practices in science. Would

5. This is not a straightforward distinction. Winsberg points out that the distinction invoked by Frigg and Reiss seems to presuppose that the choice of a model and the choice of a method to solve that model are independent of each other. This is not true.

6. Popularity that is well supported by a lot of the scholarship on computer simulation.

7. The literature on modelling is well aware of this. What is unclear is the degree to which computer simulation modelling differs from conventional modelling in this aspect.

8. This is because the code behind a lot of contemporary programs is too long and complicated for a single person to verify.

9. Similar to how all idealized models has certain "false" assumptions that do not correspond to the real system of interest.

10. This is evident by the popularity of the phrase "in-silico" experiments.

this necessarily imply that the use of computer simulation models has consequences for philosophy of science? Phrased more simply: Has the practice of using computer simulations fundamentally changed the scientific method, or has it merely provided a more efficient means of implementing the standard method(s) of science.[11]

Humphreys (2004) posits that computer simulation models challenge an anthropocentric form of empiricism that is widely held by philosophers of science. He further claims that scientific epistemology might no longer be synonymous with human epistemology, as stated best by the section in Humphreys (2004) titled 'Science Neither by the People, nor for the People' (Humphreys 2004, 6).

Humphreys also claims that computer simulation problematizes the syntactic and semantic views on theory (Parker 2014). The problem of *epistemic opacity*[12]—which refers to the fact that we don't have direct access to all the relevant steps that were taken by the computer simulation model in the process of converting input data into output results—challenges the syntactical view on theories. Computer simulation models utilizes steps that cannot be understood by using direct theoretical inferences; this is something that is not allowed by a syntactical view on theory. Moreover, Humphreys posits that computer simulation models also causes trouble for the semantic view of theory. This is due to the fact that often in computer simulation modelling, the specific representation of the theory in its mathematical or computational form is crucial to its solvability. This is something that is not allowed by a semantic view on theory.

This raises some questions: Do these tools have any real understanding of the system, or are they simply performing blind mechanical calculations? In this sense, how is a computer simulation model different from a calculator? The tasks that are performed by a computer are often described with the adjectives of being mindless and mechanical—a mere number-crunching exercise. This characterization, laden with a subtle anthropocentric value judgment, implies that all that a computer can ever do is blindly follow a sequence of simple steps that merely serves in expediting the more mundane aspects of a scientist's work.

While computer simulation models do engage in calculation and solely use preprogrammed algorithmic steps, the very process of representing a target system of interest as a model, and its articulation as an algorithm transforms the problem into a particular format. This entire process opens up new avenues for finding solutions which is something that a calculator cannot do. Nersessian (2022) substantiates this claim by noting the various affordances that simulation models provide to their creators. These are the opportunities for making connections and generating insights that would remain elusive in the absence of these simulation models (Nersessian 2022).

11. Similar questions arise whenever science incorporates new methodologies into its toolkit. More recently, similar questions were raised when scientists began to widely adopt big data algorithms in their practice (Andrews 2024a; Leonelli 2020).

12. Please refer to chapter 2 for a detailed exposition on the concept.

More often that not, computer simulation models cannot be placed in the same category as other computational tools like pocket calculators. Winsberg (2022) states that:

> ...a computer simulation is a program that is run on a computer, and that uses step-by-step methods to *explore* the approximate behavior of a mathematical model... (Winsberg 2022, emphasis added)

Note the use of the word "explore" in this technical definition implying certain notions of discovery rather than perfect determination. We wouldn't say that a pocket calculator is *exploring* a solution space when we input a simple mathematical operation like $2+2$ into it. If computer simulations were merely glorified calculators, we wouldn't describe them as exploring their model space either. This cannot merely be put down as an inconsistent use of language.

However, we should also be careful as to not make the mistake of claiming that we can categorize computer simulations as intelligent entities just because they can successfully perform a set of tasks in some narrow contexts (Burkholder 2000).[13] Computer programs outperform humans in specific tasks like playing a game of Chess or Go. These games were once thought to require creativity and were believed to be beyond the capability of mere number-crunching machines. However, advancements in computer programming have changed this perspective and led scholars to adopt a more nuanced definition of intelligence (Bringsjord and Govindarajulu 2024).

A unique perspective by Burkholder (2000) refers to computer science as an "empirical engineering science" that conducts in-silico experiments and subsequently develops theories to explain the results of these experiments. She describes a computer program that determines the average time required to retrieve a piece of data from a specific type of data structure. This involves collecting different instances of the data structure, calculating the time it takes to retrieve a data key in each of the particular instances of the data structure, and obtaining the final result by averaging the processing time overall the instances in a collection. Subsequently, she draws parallels between this particular "experiment" with other experiments in science.

However, it's unclear that this specific example has any utility beyond its immediate context of creation. If we want to study algorithms and determine the time it would take to run them on a computer, we have no choice but to use the computer as our "experimental" setup. However, in most of our computer simulation models, what matters most is whether the model results correspond to empirical data, rather than the properties of the model *algorithms* themselves. Burkholder (2000) defines an algorithm as:

13. I will return to criterion of intelligence and the basis on which we consider a computer program as being intelligent in Chapter 4.

> . . . a procedure for correctly calculating the values of a function or solving a class of problems that can be executed in a finite time and mechanically—that is, without the exercise of intelligence or ingenuity or creativity. (Burkholder 2000)

However, it's unclear as to how we should understand words like "mechanically," "ingenuity" and "creativity." This is a pressing question in light of the *epistemic opacity*[14] resulting from increasingly complex algorithms in computer simulation and the advent of black-box algorithms in AI-ML.[15]

2.6 Epistemic opacity in computer simulations

Despite debate around philosophical novelty, Parker (2014) notes how large parts of computer simulation modelling have a continuity with conventional modelling practices. The problem of external validity that was relevant for conventional modelling practices is equally relevant for computer simulations. Confirmation theory—which involves assessing support for theoretical predictions using experimental results—would similarly have a parallel formalism in computer simulations.

However, some philosophers posit that despite this continuity, the methodology of computer simulation has novel features that are unexplained by conventional philosophical frameworks. One such feature is that of *epistemic opacity*.[16] This arises due to the fact that no human can access every individual computational step of the computer simulation model simply due to the sheer number of calculations in any program. Additionally, parts of the program are epistemically opaque because we cannot directly relate the computational code to our understanding of the system (Parker 2014).

Computer simulations have inherited all of the epistemological concerns of computer programs in general. The proof of the 4-color theorem—aided as it was by a computer program—is one such example of the epistemological concerns surrounding the use of computer programs in research (Strogatz 2007). Although the construction of this program required human ingenuity, the actual verification process for the theorem's proof consisted of calculations that were so lengthy and complex that no human could verify each individual step of the proof.

The significance of the concept of epistemic opacity can also be illustrated within the method of parameterization. In the physical sciences, it's often the case that we have to model a target system using a large set of differential equations. However, some of these equations might not have exact analytical solutions. Parameterization simplifies

14. Please refer to chapter 3.
15. Taken up in chapter 4.
16. I will provide a thorough treatment of the concept of epistemic opacity in chapter 3.

these sets of differential equations by combining various variables of interest into a single parameter (Winsberg 2022). While the parameter is selected (or created) in such a way that it greatly aids the study of the phenomenon, the semantic interpretation of the parameter might have little significance in accepted theory.

For instance, intuition suggests that a good indicator of precipitation over an area is the amount of sunlight received over it. However, our simplified equations might model a system where a better parameter for predicting rainfall could be a unified quantity of a certain relationship between sunlight, dust, albedo, and other variables of interest in the model, such as:

$Precipitation\,parameter = \frac{\text{sunlight} \times \text{dust}}{\text{albedo}}$.

It is difficult to form an intuitive understand of such complicated parameters from first principles of theory. Moreover, they raise challenging metaphysical questions like determining if these parameters actually represent some phenomenon in the target system or if they are merely an artifact of the model. Sometimes, the parameter cannot even be reduced to physically comprehensible variables like sunlight, dust, and so on.

Besides parameterization, the practice of idealization poses similar philosophical questions for computer simulations as it does in conventional modelling. Because of their limited computational power, climate models of the atmosphere might idealize thousands of kilometers of the Earth's atmosphere into just tens of distinctive layers. It's also common to work with models that would idealize an entire area of thousands of square kilometers of ocean as one homogeneous entity with no internal dynamics. Subsequently, when these idealized models accurately predict long-term ocean currents and large-scale cyclone evolutions, there is confusion as to the metaphysical status of the entities that were introduced in the model. Another noteworthy feature of simulation models is that a larger model might be composed of multiple smaller models, where these smaller models might have pairwise inconsistencies among themselves. Despite this, these smaller models combine to produce a "motley model" that confirms with expected results at the larger scale (Winsberg 2022).

Answers to these metaphysical questions are important for guiding scientific practice. For instance, if we were to study a statistical model of two miscible gasses in a room, there would be a lot of strange microstates that correspond to a particular macrostate of the system. Some possible configurations include one in which all the molecules of these two gasses will perfectly separate into two parts of the room. Without being provided with the specific context, it is unclear to me if experts can make claims as to whether these minority state can be actualized in the real world. These states could very well be fictional representation that do not correspond to a real state in the target system. However, at least in some cases, these microstates do happen to correspond to real physical actualizations of the elements of the system.[17] Further

17. One such example can be seen in Perera (2015, 4:46-6:26).

investigations on this front can lead to novel insights into the theory itself.

Barberousse (2018) elaborates on some other problems in the epistemology of computer simulation concerning epistemic opacity. One such problem is the computer's lack of understanding of the differential equations on which the model is built and the user's misunderstanding of the direct outputs of the computer. This is because the computer can only "understand" 0s and 1s, and the user cannot directly make any inferences from results that are expressed in the binary language. The computer processes large quantities of data that must be processed in such a way that it is comprehensible to humans.

2.7 Epistemic opacity and black box algorithms

Epistemic opacity of computer simulations is closely related to the black-box problem in AI-ML algorithms. Considering the overlap between computer simulation algorithms and those of AI-ML, it would be beneficial to use the philosophical frameworks built to understand computer simulations and address some of the epistemological questions that arise for these novel AI-ML algorithms.[18]

Subsequent computer simulation models may increasingly incorporate AI-ML algorithms and other such novel algorithms, increasing their epistemic opacity for human experts. This raises concerns of whether we are warranted in believing the results of computer programs that increasingly rely on algorithms that we cannot directly assess or understand. Alongside epistemological issues, epistemic opacity poses various ethical questions as well. This is because many decisions like making policies on climate action and performing medical diagnoses are now relying on the results of these programs. In the next chapter, I will explicate the concept of epistemic opacity in detail and motivate the need to study its implications for scientific practice.

18. This is something that is already being done by people working on AI-ML algorithms like Juan Manuel Durán and Jongsma (2021)

CHAPTER 3

EPISTEMIC OPACTIY

3.1 Introduction

Epistemic opacity is a form of epistemic inaccessibility first introduced by Humphreys (2004). Humphreys claimed that computer programs[1] are inherently inaccessible to humans due to the limitations of our cognition. A simple computer program might comprise millions of computational steps converting input data into output values. The sheer magnitude of these steps makes the program opaque to the human cognizer (also referred to as the agent). Epistemic opacity is a conditional property of the relationship between the computer program and the agent trying to access the program.[2]

Scientists and philosophers are often confronted with models that can accurately predict phenomena but fail to provide any explanations for the same. Epistemic opacity (henceforth opacity) is situated in another larger question in general epistemology concerning the warrants for accepting the outcomes of an epistemic process when we do not have a clear understanding of the epistemological process itself. The process can generate claims like the output of a particular run of a computer simulation model, or even the measured value from an experimental setup.

An adequate response to the skeptical challenge of opacity will need to situate itself within the broader issue of transparent and opaque knowledge in epistemology. There is no doubt that computer simulation represent a new methodology for science. However, it isn't necessarily the case that they pose novel issues for contemporary frameworks in philosophy of science.

Many of these computer programs, like computer simulation models, are widely employed in various domains of science. Humphreys (2004) posits that conventional

1. Humphreys primarily talks about computer simulations. However, the argument can be extended for computer programs in general.

2. Humphreys (2004) defines a program to be opaque if the agent does not have access to the epistemically relevant elements of the program.

theories of justification in philosophy of science cannot account for this novel methodology. Philosophers have justified conclusions by tracing them back to accepted premises using step-wise logical inferences. However, it is impossible to trace the results of many computer programs to their input data or to justify the outcomes of the program using accepted theory (Humphreys 2004). Nevertheless, the predictive success of these programs is undeniable. This has prompted philosophers to formulate novel conceptual frameworks to justify the use of computer programs in science.

Contrary to the general problem of epistemic inaccessibility of various phenomena by human agents, I will constrain the usage of epistemic opacity to the domain of computer programs to avoid trivializing the concept. Limiting our scope in this manner aligns with recent literature explicating the concept of opacity. San Pedro (2020) restricts his usage of the concept of opacity to refer to "computer processes designed, built, run, and so on, by human agents." Another conditional property that can be added is to only refer to agents who are experts in the domain. Even for experts, any particular program would be rendered transparent only after a certain amount of training and cognitive development.[3] A 3-year-old will not have any understanding of a simple Lotka–Volterra prey-predator simulation model. However, the same model will be transparent to a trained theoretical ecologist. Therefore, I will limit my scope by restricting the usage of opacity to talk about agents that are trained experts in their domains.

In the following sections, I will introduce Humphreys (2004)'s foundational formulation of the concept of opacity in computer simulations. Subsequently, I will critique various aspects of Humphreys' framework and review arguments from the philosophical literature about the epistemic desiderata of computer simulations.

3.2 Humphreys' foundational definition of opacity

A computer simulation represents real-world phenomena by defining a set of variables that interact to determine the system's state. The simulation then examines how this state evolves based on certain predefined rules. In the physical sciences, a computer simulation can be employed to model the evolving state of a system by a multitude of successive steps. Models commonly used in science might execute more than a million such steps as the system evolves. Because no human can manually verify each and every one of these individual steps, Humphreys (2004) argues that the results of a computer simulation model have to be accepted at face value. He further states that the inability to access all the relevant steps of the evolving system leads to a novel

3. Alvarado has similar views: "One can, for example, imagine that at every point in someone's lives some things will be epistemically opaque without representing a significant epistemic challenge. This makes the concept seem a little trivial. In other words, almost every single process a growing human will come to understand was at some point epistemically opaque to them" (Alvarado 2021, 3).

epistemological issue in philosophy of science.

In Humphreys (2009), Humphreys builds on his earlier work and modifies the definition of epistemic opacity to generalize the concept beyond computer simulations. He introduces the idea of the epistemic opacity of a particular process P for a cognitive agent X, where X does not have access to all the steps that justify P.

In this manner, even an NMR machine can be epistemically opaque to an experimenter if she does not have access to all the "epistemically relevant elements" needed to justify the results of the NMR. Justification and understanding of the results of the NMR machine could include knowledge of how the NMR machine was constructed, the experimental support for the validity of NMR results and so on. However, scientists seem to justifiably draw inferences from instruments even when they do not have a perfect understanding of all the details of the instrument (Beisbart 2021).

As Humphreys does not propose any direct solutions to address opacity, Juan M. Durán and Formanek (2018) provide an alternate framework wherein they explore the possibility of acknowledging opacity in computer simulations without losing trust in the results of computer simulations. Durán and Formanek (2018, 653) modify the epistemic framework of *process reliabilism* and extend its scope to computer simulations. They call this new framework *computational reliabilism*. The principal idea behind reliabilism seeks to warrant epistemic justification for accepting the results of a process without having a complete understanding of the process itself.

Opacity seems to be intimately linked to the agent's *understanding* (or lack thereof) of the computer simulation. One way of conceptualizing *understanding* a computer simulation is possessing knowledge about the computer simulation that extends beyond the knowledge of individual simulation results. This will allow the agent to perform counterfactual reasoning and explain the outcomes of a simulation model based on a particular set of input data.

Beisbart (2021) uses this idea of understanding to problematize Humphreys (2009)'s framework by introducing the concept of a "superhuman cognitive agent" that has access to all the steps of a computational process. According to Humphreys' definition, the computer simulation model will be epistemically transparent to this superhuman agent. However, even while having access to all the individual steps of the computational process, the agent might not have an *understanding* of the computer simulation. This is because other important epistemic properties might continue to be elusive to the superhuman cognitive agent. One of these epistemic properties is the inability to understand the connections between different individual steps of the computer simulation. The agent might know that a variable is exponentially increasing with each step; however, it might not know what is causing the variable to increase and whether it is affecting any other variables in the computer simulation. This undermines the utility of Humphreys (2009)'s foundational framework, leading Beisbart (2021) to

suggest modifications to the definition of opacity. Having introduced Humphreys seminal contribution to the concept, I will now draw from more recent work to define the concept in a way that is suitable for this thesis.

In responding to Humphreys (2004, 2009) works philosophers have refined the concept of opacity. Having acknowledged Humphreys' seminal contribution to the concept of opacity, I will now draw from recent scholarship to refine the definition of opacity in a way that is most helpful for this thesis.

3.3 Critiquing Humphreys' framework

3.3.1 Different notions of opacity

A computer simulation can be described at various levels: the physical, computational, and representational level (Beisbart 2021). Humphreys conflates the representational level of a computer simulation with its computational level, however, it is possible that a system that is opaque at one level, might not be so at another level.

In contrast to Humphreys' definition of opacity, Imbert (2017, as cited in Beisbart 2021) proposes a notion of opacity that arises due to the collaboration of various experts from different fields in the process of constructing a computer simulation (and not due to the great number of computational steps). This leads to a situation where no single expert has a complete understanding of all aspects of the computer simulation. However, it is possible that this community of experts as a whole can be said to have epistemic transparency towards the computer simulation. This form of *social opacity* gives a complementary perspective in understanding other technical elements of a computer simulation, like modules, libraries, and other such code that is written by other programmers. These sets of code are frequently unverifiable by the creator of the computer simulation (Juan M. Durán 2018, 106).

In addition to social opacity, Durán describes *technological opacity* that arises from researchers' lack of a deeper understanding of the instruments employed in their scientific activities. A more relevant form of opacity for computer simulation models is *mathematical opacity*. Internalist forms of mathematical opacity highlight how a researcher might be unable to completely understand the computations of a computer simulation due to the sheer complexity of the mathematical model that is being employed. Externalist forms of mathematical opacity describe the reliance of a researcher on a computer simulation model to solve complex mathematical models. The externalist form of opacity arises because the researcher cannot verify the results of the computer simulation model by her own independent means (Durán 2018).

3.3.2 Epistemically relevant elements in a computer simulation

Humphreys (2004)'s definition of opacity centers around the concept of "epistemically relevant elements" in a computer simulation. But what determines epistemic relevancy? Durán (2018, 104) introduces a technical definition of an *epistemically relevant element* in a computational process as any component—like a function or a variable—that is involved in the computation of the model for the purpose of rendering results.

Durán and Formanek (2018) also propose a more general conceptualization of the *epistemic elements* as "steps of justification" in an argument. An agent has knowledge of these steps if and only if the steps fulfill the criteria of surveyability and accessibility for the agent's cognition. In doing so, they modify Humphreys' definition of opacity for a cognitive agent X. In the new definition, a process is epistemically opaque to X, if X can't survey, and does not have access to all the steps of the justification for the epistemic process.

Durán and Formanek (2018) present an example of a "hello world" program and argue that the successful rendering of the result of the program is not epistemically opaque to a programmer. This is because the programmer can understand the results of the program based on his previous knowledge of the CPU, the programming language, the installed libraries of code and so on.

The skeptic can argue that whatever warrants of justification that we can provide for accepting the results of a computer simulation are inadequate. In response, Durán and Formanek (2018) note the prudency in engaging with the commonsensical meaning of justification in terms of the set of tasks a human can perform in her natural biological timescale. Any demands for justifications should be constrained by the limits of human biology in space, time, and cognitive capacity (Durán and Formanek 2018).

3.3.3 Degrees of opacity

San Pedro (2020) criticizes attempts to view epistemic opacity as a "yes-no" property of a process. Instead, they advocate for viewing opacity as an essentially graded property. This graded concept will go beyond a qualitative "yes-no" qualification and will help connect computer simulation to other modelling practices. If we can have various degrees of understanding for a computer simulation, we can also have degrees of opacity.

However, accepting the spectrum conceptualization of opacity raises questions about the utility of this concept beyond the truism that all computer simulation have some degree of opacity. In response, San Pedro (2020) advocates for further work in determining factors that can help scientists quantitatively measure the degree of opacity of a computer simulation and explore methods by which we can make a computer simulation more transparent to epistemological considerations. This can be

done by combining computer simulation with other modelling practices or by direct comparison with empirical data. In this manner, a computer simulation will be labeled as epistemically opaque if it is below a minimum opacity threshold.

Going back to the example of the "hello world" program, it is true that a programmer with more knowledge of the hardware of the CPU will view the program as being less opaque compared to a programmer who is not trained in electronics. Nevertheless, the program remains epistemically transparent for both of them. Although individual humans will always differ in their abilities, the principal interest is in defining a minimum threshold of access. One way to do this is to establish a threshold of opacity for the abilities of an average scientist working with the computer simulation model (Durán and Formanek 2018). However, this might vary depending on the epistemic goals of the computer simulation.

3.4 Epistemic goals of an opaque computer simulation

There are various epistemic goals of a computer simulation. Some of these are prediction, understanding, heuristic explorations, and so on. I will now try and demonstrate how the interactions between these goals have consequences for the conceptualization of opacity. This will highlight the importance of reviewing the epistemic goals of explanation, discovery, and prediction in more detail which I will revisit in chapter 5.

3.4.1 Prediction

Humphreys argues that opacity results from the practical constraints of the inability of humans to slow down the computational process and verify its stepwise computations (Humphreys 2004, 150). A prediction, as opposed to a simple claim or inference, has to conclude before the state of affairs it is predicting.[4] Therefore, the question of the computational time of running a computer simulation goes beyond considerations of engineering and efficiency and is directly relevant to philosophy (Humphreys 2009).

There are other epistemic goals of a computer simulation other than the obvious goal of making predictions. Beisbart (2021) notes that most simulations are constructed in order to achieve particular context-specific outcomes. It is possible that the only way to understand computer simulations is by situating each computer simulation within their specific context of application and being cognizant of their particular epistemic goals. There could exist cases where computer simulations need to only make predictions without requiring to provide the agent with any theoretical understanding. However, we can raise our epistemic expectations by demanding that computer simulation models

4. I will return to the central importance given to the notion of temporality in predictions in the later chapters.

should also provide explanatory value and theoretical understanding to the agent. This will have consequences for how a computer simulation model is constructed, trained, and evaluated.[5]

3.4.2 Understanding

Beyond their predictive capabilities, scientists also use computer simulation to gain a better understanding of a system. Identifying the underlying patterns in the predictions of a simulation might provide leads for further research. However, epistemic opacity presents a formidable challenge against the epistemic goal of understanding.

There is significant divergence in the popular accounts of *understanding* in epistemology. However, all accounts seem to converge on the fact that there can be various forms of understanding and that there are degrees to understanding (Lipton 2001, as cited in Durán 2018).[6] Suppose a lack of understanding on the part of the agent was the principal cause for the epistemic opacity of a computer simulation. In that case, the property of understanding being a graded concept would align with my treatment of opacity as a graded concept itself.

Beisbart capitalizes on these insights and modifies Humphreys' foundational definition of opacity by shifting emphasis from individual (computational) steps of justification to a lack of understanding about the epistemically opaque process. In this formalism, some X is epistemically opaque if, and only if, it is difficult, if not impossible, to know and to understand why the outcomes of X arise (Beisbart 2021, 11659). For Beisbart, understanding of a proposition P entails a dual requirement of:

1. Knowledge of why P, and

2. understanding why P,

where P can be the outcome of a computer simulation (Beisbart 2021). This seems correct, because the ability of an agent to understand why P was the resulting output for a particular input is a good indicator of the agent's understanding of a computer simulation. The agent can gain understanding of a computer simulation by specifying the primary goal towards which the computer simulation is being implemented and by acquiring abilities of counterfactual reasoning on the results of the computer simulation. Scientists can also gain understanding of the computer simulation by running multiple instances of the simulation where they change variables and parameters to study the dynamics of a model.[7]

5. I will return to the question of the possibility of separating these epistemic goals in the later chapters.

6. Moreover, note that the understanding of a process can be conceptualized as having access to particular kinds of knowledge about the process (Lipton 2001 as cited in Durán 2018, 118).

7. This might not be a sufficient condition in practice. Often, a model might take months of compu-

3.4.3 Heuristic purposes

Computer simulations are widely employed in successfully exploring promising avenues of pursuit in research, which might qualify them as a heuristic tool that can be employed in the context of discovery. In this manner, opacity need not diminish the capacity of computer programs like computer simulation to lead scientists to significant and justifiable breakthroughs (Duede 2023, 1089).[8]

Scientists also use computer simulations to improve their personal understanding of the phenomenon being modeled. This complicates the notion of opacity as a relation between a computer simulation model and a human unassisted by any machine (Beisbart 2021). San Pedro (2020) highlights the fundamental link between the concept of opacity and the particular cognitive agents for whom the process is opaque. Any question about opacity will raise the question of "opacity for who?" and even if the answer were "humans," what kind of humans? Maybe the computer simulation model is opaque for a new intern who joins a research lab; however, closely working with the computer simulation for years might make the model transparent to the intern. Humans can use other machines, computational programs, and even other computer simulations to understand their own computer simulation model.

Perhaps the correct cognitive agent with respect to which we need to define opacity is a hybrid entity of human-computer-simulation, for instance, a human aided by the affordances of a computer simulation like graphs and diagrams. In this context, humans can take the aid of computer programs like graphical representations to understand complex data and mathematical formalisms. If this were the case, it would support Humphreys' argument that science is moving beyond an anthropocentric epistemology (Humphreys 2004).

It is worthwhile to investigate trends in the construction of computer simulation models that are founded on increasingly opaque algorithms. This might imply that computer simulation models are being constructed to achieve the principal epistemic goal of heuristic explorations (or predictions), which can come at the cost of the epistemic goal of gaining theoretical understanding.[9]

tational time to complete a single run on a particular set of values. This shatters any hope of applying counterfactual reasoning, which would require data collected from multiple runs of the model (Beisbart 2021).

8. Duede is talking about machine learning models, however, the argument can easily be extended to computer simulations.

9. I will return to the epistemic trade-offs in modelling exercises in the chapter 7.

3.5 Justification for using opaque computer simulations in science

Should scientists take care in reducing opacity in computer simulations, or should opacity be accepted as a fundamental property of computer simulations?

Humphreys (2004) acknowledges that our inability to access all the details of the model need not be a sufficient condition for loss of knowledge and understanding. He illustrates this point by discussing modelling practices like idealization, which while limiting the access to the model's description, actually allows experts to gain *more* knowledge than was otherwise possible. For example, a statistical model of a gas can help us gain more knowledge and understanding of the macroscopic description of temperature, pressure, and so on by ignoring microscopic interactions of individual molecules (Humphreys 2004).

Some philosophers take this to mean that a good computer simulation (like a correct idealization or abstraction) can, despite opacity, actually increase our trust in its results (Durán 2018, Beisbart 2021). It might seem unreasonable to use a computer simulation to explain a phenomenon when we do not entirely understand the computer simulation itself. However, we frequently use this very approach for attempting various kinds of explanations. An explanation can be conceptualized as having access to the right kind of knowledge. For a proposition P; understanding "why P" often involves gaining more knowledge about the explanatory elements of the processes that led to P, rather than a perfect description of P itself.[10]

Humphreys (2004; 2009) does not recognize any direct method by which computer simulations could be made more transparent and instead advocates for indirect methods that can warrant accepting the results of computer simulations. This includes steps like verifying the outputs with empirical data, cross-checking simulation models with ideal solutions in known cases, perturbing input data, and so on.

However, Duede (2023) argues that philosophers might be unreasonably skeptical about the epistemological foundations of these tools. What we need is to distinguish between good and bad computer simulation practices without resorting to a radical skepticism that will sweepingly generalize all computer simulation models as being unreliable due to their opacity.

Durán (2018) draws a helpful conceptual distinction between opacity and other sources of computational errors. While considering computational errors, good programming practice and various other verification and validation methods can help recapture trust in the results of a computer simulation. However, opacity is a deeper

10. This is a necessary step to avoid the issue of the infinite "why-regress", where an agent can keep on demanding another explanation for whatever explanation we provide ad infinitum (Lipton 2001, as cited in Durán 2018, 118).

problem that challenges the foundations of any knowledge claim that can be derived from a computer simulation, prompting Durán to refer to opacity as a "permanent loss of knowledge" and an "irreversible uncertainty." Framed in this way, opacity might be a position of radical epistemological skepticism that is unamenable to any intervention.

In order to challenge the skeptical argument against computer simulation results, Durán and Formanek (2018) resort to external sources of validation for computer simulations. They present an account of *computational reliabilism*, which says that accepting the results of a computer simulation is warranted on certain reliability properties of the computer simulation model itself.[11] Scientists need to provide independent reasons to justify the results of a computer simulation. As an example, the authors of the mathematical proof for the four colour theorem had to provide independent reasons to support the reliability of the program that was used in their proof (Durán 2018, p. 105).

Durán and Formanek (2018) provide four sources for attributing reliability to a computational process. These four sources are:

1. Verification and validation methods
2. Robustness analysis for computer simulations
3. A history of (un)successful implementations
4. Expert knowledge

I will illustrate the first and last of these points. Verification and validation methods can add reliability to the results of a computer simulation by ensuring that the computer simulation will be constructed on the basis of established theory and that its results will match with empirical data. When computer simulations are constructed in close contact with experts, their judgement and experience are incorporated into the computer simulation model.

3.6 How opacity undermines epistemic goals

Science does not solely aim towards predicting data. It also aims to explain the natural phenomena that give rise to the data. Opacity hinders the realization of this epistemic goal of explanation. Besibart (2021) identifies the cause of opacity as a lack of understanding on the part of the agent interacting with the computer program's outcomes.

11. Computational reliabilism is a modified form of the justificatory account of process reliabilism applied on computer programs. Process reliabilism warrants the acceptance of the results of a process by demonstrating the reliability of the process. If it can be proved that a process has a history of accurately producing knowledge in the past, we can accept whatever result it is producing in the present because the *process* of knowledge generation has been shown to be reliable.

He further notes that one can make the program more transparent by gaining a better understanding of it.[12]

Before proceeding further, one has to clarify how an agent can ever understand an epistemic process that is opaque. As I have already mentioned in the previous sections, this is a tricky question as defining the concept of understanding itself has been a challenge in epistemology for centuries. Nevertheless, for my purposes, I only need to point out that like opacity, understanding is also a graded and agent-relative concept. Different people can have a better or worse understanding of a computer program, implying that the program is more or less transparent to these different agents. Treating understanding and opacity as graded concepts allows us to think of ways in which they can be quantified in specific domains and help plan interventions that can change how they are expressed to the agents (San Pedro 2020; Beisbart 2021).

I should note the possibility that opacity of the form that is exhibited by computer programs is not a novel phenomenon. If this were true, researchers could simply reduce opacity using standard scientific practice widely employed in scientific modelling. For instance, San Pedro (2020) describes the model of a simple pendulum and argues that approximating the infinite $sin(x)$ expansion for small angles renders the simple harmonic model of the pendulum transparent to us. The essential takeaway is that the initial problem of using numerical solutions to approximate the $sin(x)$ function in an opaque simulation has been circumvented by employing approximations and other mathematical techniques that are a part of established theory.

However, it is not always possible to use established theory to understand a computer program. In the next chapter, I will introduce new methodologies that employ AI techniques to scientific modelling, and explore the challenges of interpreting their results through the lens of established theory.

12. One way to know that we have gained a better understanding of a program is if we can perform counterfactual reasoning on its space of input-output records.

CHAPTER 4

EPISTEMOLOGY OF MACHINE LEARNING IN SCIENCE

AI (Artificial Intelligence) is currently being used as an umbrella term for a range of systems and phenomena. AI can be an emergent property of a swarm, AI can be a large systematic collection of logical rules, AI can be a novel big data algorithm being used by scientists, and the list goes on. We are currently in the paradigm of connectionist AI as evidenced by the rapid development of connectionist architectures like large language models (LLMs, like ChatGPT), other popular machine learning algorithms, and so on. In the field of AI, Machine Learning (ML) is a specific technique comprising a set of algorithms that are used to train a computer model on data-sets to discover novel patterns. However, why should we refer to ML as an instance of AI in the first place? What are the differences between an ML model as opposed to any other mathematical model that is typically used in science?

Let alone AI, some scholars go so far as to suggest that even ML as a word is being used as a referent for such a broad range of systems that any two instances taken for comparison that are referred to as "ML" will very little meaningful similarities (Andrews 2024a). Despite this caveat about an overly broad definition, I think what is meaningfully common to all cases of ML is their categorization under the umbrella term of AI.

Who decides if a particular computer program deserves to be termed as an instance of AI? By some definitions in the philosophy of AI, no system that has yet been developed can be termed intelligent (Bringsjord and Govindarajulu 2024). On the other hand, we already seem to have widely accessible models that have passed the Turing test (Turney 2024; Biever 2023). It will be illuminating to review how and why these specific sets of techniques that comprise ML are called AI, the appropriateness of the label, and the implications of using this label for scientists' conceptualization of these models. It

is possible that because ML models are put under the category of AI[1], scholars seem to automatically assign a greater degree of agency and autonomy to ML models as compared to conventional computer simulations.

At this point, I will define my usage of the fundamental concept of *learning* such that it can capture instances of both human learning and machine learning. *Learning* is an internal *change* in an entity such that the behaviour of the entity after the change is different from the behaviour before the change.[2] I will now build on this definition to posit that ML is an instance of AI because unlike pocket calculators, ML is a system that *learns*. An ML model has the ability to change its own parameters in order to minimize a pre-programmed loss function, and the capability of the model to do so can also justify ascribing a degree of *agency* to an ML model.

Some critics contest that contemporary AI-ML models are *just* a set of functions and parameters and do not have any form of agency, or even intelligence. However, even if AI-ML models are just functions and parameters trained on a specific set of data that were programmed to realize very narrow and specific goals, these models can nevertheless demonstrate emergent properties that can go beyond the data or the immediate training objectives that they had.

4.1 The new methodology of AI in science

In the philosophical literature, AI refers to attempts at creating intelligent animals (Bringsjord and Govindarajulu 2024). For my purposes, I will focus on AI as it is used to refer to a set of computational techniques like machine learning, deep learning, and so on that are being employed in various domains of science (among other fields of study).

As I have already noted before, there are various forms of epistemic opacity which can make it difficult for scientists to evaluate the outputs of a model that has been implemented using a computer (Beisbart 2021). The use of increasingly complex computer models comprising AI algorithms in various domains of science has exaggerated this problem. These AI models are trained on large data sets to help researchers discern patterns within the data and make novel predictions. This data-driven approach toward science prides itself on being free from human biases and the subjectivity associated with it (Srećković, Berber, and Filipović 2022).

Facchini and Termine (2022) note that studies of highly complex phenomena are widely employing AI tools. Moreover, the use of these tools is being presented as a novel methodology in science.[3] The authors present a working definition of the

1. It is entirely possible that the arrow of causality runs in precisely the other way.
2. Thanks to K.P. Mohanan for highlighting the need for this definition.
3. Humphreys (2004) made this original claim for computer simulations. Similar claims about ML are

standard methodology in science[4] as the formulation and experimental evaluation of hypothesis with the goal of explaining observable facts. They subsequently claim that AI methodologies do not cleanly fit into this model of doing science.[5]

Andrews (2024a) provides us with a helpful categorizing scheme to organize a wide range of claims about AI models. The philosophical novelty of the use of AI models in science rests on the assumption that these models represent a fundamentally different methodology of doing science. This *distinction claim* is subsequently used as evidence for the *disruption claim*, which states that the widespread use of AI models will cause a fissure in the continuity of scientific practice. One such disruption could be caused by scientists relying entirely on theory-agnostic, data-driven methods of predictions which will make anthropocentric epistemic desiderata like theoretical understanding and explanatory value obsolete (Andrews 2024a).

Although these AI models are very powerful in terms of their predictive abilities, they provide little understanding of the causes that led to their predictions. They also fail to give us accurate knowledge of how we might successfully intervene on the system of interest. The inability to provide an understanding of the target phenomenon to human agents is a particular form of opacity.

4.2 The philosophical novelty of AI models

In a special issue journal article titled 'Machine Learning: Prediction Without Explanation?', F. J. Boge, Grünke, and Hillerbrand (2022) posit that ML[6] is making a fundamental shift in our philosophical understanding of the concept of explanation in science. These ideas are corroborated by Srećković et al. (2022) when they posit that machine learning (and other AI techniques) will sever the functional relationship between explanation and prediction. The practice of scientists using explanations in order to make novel predictions about the future will become obsolete. Srećković et al. posits that AI models do not rely on an anthropocentric methodology of doing science and can directly infer predictions from raw data using theory-agnostic modelling methods.

These sets of arguments seem to presume that as these AI models improve with time, they can be adopted into various other domains of science. Subsequently, the AI models will be able to make novel and accurate predictions, unhindered by human subjectivity and cognitive limitations, and will end up out-competing human scientists

coupled with claims of methodological novelty of big data algorithms in science (Leonelli 2020).

4. I will bracket the discussion on the plurality of methods that fall under the "scientific method", and the difficulties associated in classifying the same.

5. There is an ongoing debate about the possibility of reducing all scientific activity to one single unique scientific methodology without acknowledging the role of human creativity and imagination in the discovery process.

6. Which is a subset of various techniques in connectionist AI.

who have to rely on mere expertise and theoretical understanding in order to advance science.

Andrews (2024) responds to Srećković et al.'s arguments about the *distinctness* claim by invoking the concept of theory-ladenness in the data that the AI models are trained on. This claim is correct, and challenges the theory-agnostic conception of modelling that is being advocated by Srećković et al. (2022).

However, I would also like to point out that the interpretation of the results of AI models is not theory-agnostic. Scientists do not accept all the outputs of an AI model, and the deliberation on which outputs of a model are to be accepted or discarded is provided by the theoretical understanding of the domain expert.[7] The fact that there are human scientists at the end of the modelling pipeline who will eventually interpret and integrate the results of the models will limit the disruptive potential of this shift.

I will now analyze the disruptive and distinctive claims by Srećković, Berber, and Filipović (2022) in more detail.

4.3 AI-ML models making predictions without explanations

Srećković, Berber, and Filipović (2022) state that AI-ML models make predictions without explanations. The authors describe a functional relationship between predictions and explanations wherein explanations are useful to the extent that they can help scientists make more predictions.[8] They further claim that these ML models represent a novel "theory-agnostic" method of conducting science.

The larger debate is about whether these ML models are in fact outperforming (or will inevitably outperform) the theoretically grounded models of conventional science. Although AI models surely outperform other conventional models in making inferences from small data sets, concerns can always be raised on whether they might be overfitting the data or if they are actually representing the true causal structures behind the superficial correlations that they are discovering.

For instance, AI can sometimes identify correlations that allows it to make highly accurate predictions in the training sample, however, these correlations lack any real meaning or causal explanation. For instance, take the case of a ML binary classifier that was trained to distinguish between images of wolves and dogs. Although the learned model was able to make accurate predictions in the training data set, it was noticed that the model would apply the label of "wolves" to images of dogs in snowy backgrounds.

7. Similarly, what parts of the available data sets are to be used in training an AI model is only informed by domain expertise.

8. The authors also note that explanations can have some intrinsic value, but this value is only limited to satisfying the working scientist's "psychology".

The experts later realised that instead of learning the actual distinguishing features between wolves and dogs, the AI was relying on the background of the image to make its prediction—for the ML model, if an animal is placed in front of a snowy background, then it is a wolf![9] This problem arises because ML uncovers statistical patterns in data rather than understanding the relevant underlying concepts.[10] This is obviously a major problem that undermines epistemic trust in the result of opaque models. As Sullivan (2022) notes:

> We want some indication that the model is picking out the real difference makers for identifying a given disease and not proxies, general rules of thumb, or *artefacts* within a particular dataset. (Sullivan 2022, 21, emphasis added)

All outputs of an AI-ML model are not afforded the same epistemic status. Some results of ML models are spurious correlations while others can be used to make genuine scientific predictions. Just because an AI model is able to find a high degree of correlation in the data, does not necessarily mean that the finding is worthy of any scientific merit and is absolved of theoretical considerations.

In light of the above distinction, I will draw attention to a major assumption made by Srećković, Berber, and Filipović (2022) in the supposed independence of predictive and explanatory capabilities of a model, which has also been used to make the claim that ML models perform "predictions without explanations". In the later chapters, I will argue that it is not possible to draw such a clean seperation between the predictive and explanatory capabilities of a model.[11] Andrews (2024b) responds to these claims of theory-agnostic scientific modelling by noting that all observations are theory-laden, and ML trains on data sets that are a representation of such theory-laden observations. This implies that if AI-ML models use data that represent theory-laden observations, then by proxy, ML models also use theory. She further argues that theory will always have input at some point in the pipeline of the development of these ML models. For example, theory can be used to process the data that the ML model is trained on, to deliberate on the choice of the appropriate ML architecture, or to evaluate the results of the ML model (Andrews 2024a).

Duede (2023) defends the practice of scientists attempting to devise ad hoc explanations in order to understand the predictive success behind a particular opaque ML model and then trying to integrate these insights to advance their theory. Although

9. This example is described in Gadye (2019); and is further analyzed in Rudin (2019).

10. Uncovering spurious correlations in big data has been a rampant problem even before the popularity of ML. For instance, note the high degree of correlation between bachelor's degrees awarded in engineering and electricity generation in Cambodia (Vigen, n.d.). Despite the high degree of correlation, it is obvious that this is a coincidence and does not hint at any underlying causal structure.

11. Evidence of this can also be found in the decoherence among review articles that aimed to identify the epistemic trade-offs in ML modelling in various domains of science. Please refer to section 8.4.

there is enormous epistemic value in using ML models in this manner, we should note that any such ad hoc explanation of an opaque model is incomplete; with some scholars saying that any such explanation is always incorrect and cannot be said to accurately represent the original ML model (Rudin 2019).[12] However, other scholars respond by arguing that making an accurately similar model of the target system is not a necessary condition to construct an epistemically virtuous representation (Sullivan 2023).

Despite their differences, all scholars involved in the debate agree that ML models aid in making predictions. Another difficult arises in the process of ascertaining the epistemic status of computer-made elements that are generated by the ML model in the context of a particular modelling exercise.[13]

4.4 Data-driven parameterization as a distinctive feature of ML models

Unlike conventional computer programs, ML can create change its own parameters to minimize its loss function. AI-ML models create computer elements (like numerical functions and parameters) to achieve their programmed goals. The programmed goal refers to a simple mathematical goal, like minimizing a mathematical utility function by exploring a parameter space (Bringsjord and Govindarajulu 2024). This function can be a correlation between any set of input-output data sets. Scientists need only represent their problem in a form that can be analyzed using these ML algorithms, and they are almost guaranteed to obtain a function imbued with high predictive capabilities in similar data sets.

This is where the *distinction claim* (Andrews 2024a) about the use of AI-ML models becomes interesting. At their core, AI-ML models—like computer simulations and other computing models in their ancestry—are mathematical and statistical models. There is a disappearing line on the spectrum of mathematical complexity as we go from computer simulations to AI-ML models. This is taken as evidence to support the position that because these AI-ML models are not fundamentally different from other modelling approaches in science, they will not cause any major disruptive fissure in scientific practice.

However, I think that conceptualizing AI-ML models in this manner is an oversimplification. ML models have the unique ability to create their own parameters and model the target phenomena independent of theoretical considerations. We need more substantive philosophical conceptualization on the fundamental concepts in AI-ML

12. Rudin (2019) advocates for constructing inherently transparent models, like interpretable AI-ML models, from the get-go. This is in contrast to constructing an opaque ML and then subsequently attempting to explain its predictions in an ad hoc manner.

13. Like the correlation functions mentioned in the case study taken up by Duede (2023).

modelling (like prediction, explanation, discovery, and so on) before we can ascertain whether these models represent a distinct methodology of conducting science. I will attempt to do this in chapter 6.

4.5 The unprecedented epistemic capabilities of AI-ML models

To illustrate why AI-ML models have the potential to raise novel epistemological questions, I will entertain certain arguments which claim that the novelty of these ML models is responsible for their unprecedented epistemic capabilities.

Can an AI-ML model make scientific explanations? Answering this question would first require examining the nature of explanations that are devised by human scientists themselves. One way scientists devise explanations is by constructing unobservables like concepts, laws, and so on that can be used to understand and unify disparate phenomena (Woodward 2014). If this is a sufficient requirement for an explanation, then it is possible that ML models can devise explanations if they can construct such unobservables.

An ML model can use the training data set to create its own parameters, which can in turn be used to explain and predict more observational generalizations. Highlighting the distinct nature of the parameters of an ML model can help us draw a contrast between AI-ML models and conventional computer simulations. Most computer simulations are *downward* in their epistemology, acting top-down from theory to observations (Winsberg 2022). In contrast to this, the distinct feature of AI-ML models is that these models go *upward* from observations to observational generalizations and theory.

The fact that these parameters are not humanly understandable implies that they cannot be integrated into scientific theory; however, it is possible that these artifacts be thought of as the elements that can comprise the structure of a non-anthropocentric scientific theory that is comprehensible to AI-ML models.[14]

We can illustrate the preceding idea with an example. One reason as to why Newton's theory was accepted was due to its ability in successfully unifying phenomena as wide-ranging as the motion of falling cannon balls and revolving heavenly bodies. Similarly, a powerful AI architecture, like the LLM models at the base of ChatGPT, can find patterns in their training data which provides them the capabilities of achieving epistemic success in a wide range of phenomena that were not a part of its original training data set. In large part, it is this very ability of emergent properties of AI models

14. This is related to claims of incommensurability and underdetermination of theory by data which I will revisit in the subsequent sections (Newton-Smith 2000).

that make them an interesting object of study for scientists who are attempting to construct models that do not overfit and can generalize far beyond their training data. This predictive success is often the warrant that justifies scientists acceptance of the results of AI models.

4.6 Differentiating between outputs versus results

At this point, it will be helpful to clarify the definition of a few concepts in my present usage. Beisbart (2021) distinguishes between the *output*, the *outcome*, and the *result* of a computer simulation model. The model's *output* in the form of raw data is subsequently interpreted as the state of a system and is converted into an *outcome*. The outcomes of multiple runs of the model will finally lead to conclusive *result* for the simulation model. This framework is also helpful in understanding the difference between the outputs, outcomes, results, and predictions of an AI model.

Despite their differences, both Andrews (2024) and Srećković et al. (2022) do not explicitly distinguish between the outputs of an AI model and the final predictions or discovery claims resulting as the end product of a modelling exercise.[15] Like any other model, the process of constructing an AI model goes through a graveyard of failed configurations that had to be discarded by the expert. An AI model is fallible in the present, and its results can turn out to be wrong in the face of new evidence in the future. However, unlike conventional models, the problem of opacity complicates the present epistemic status of the outputs of an AI model. I will argue that the outputs of an AI model can only be accepted as scientific predictions when these outputs can be understood and explained by experts in the field.[16]

4.7 Opacity in AI models

Scientists have expressed concerns over the widespread use of AI tools and models. When researchers were surveyed on the negative impact of AI, more than half referred to the possibility of "…reliance on pattern recognition without understanding" as just one among various disadvantages (Van Noorden and Perkel 2023, 674). A lot of AI models aid scientists in making accurate predictions in their domains. However, can scientists justify the results of their work if these models are not epistemically transparent to them?

15. I will come back to this distinction when I respond to Duede (2023)'s work in chapter 7. This will be one of the central novel arguments of the thesis.

16. This is but one mode of justification that will warrant scientists in accepting the results of these models. I will revisit the different modes of justification in AI-ML in chapter 8.

The previous question can only be answered if we explicitly state the numerous epistemic goals of science. If the deployment of these AI tools were to maximize all of science's epistemic desiderata then there would be no debate about their implementation.[17] But it seems that AI models lead to a trade-off between some epistemic desiderata, like choosing between predictive power of the model versus the theoretical understanding of the scientist.

A human scientist has a broader understanding of a scientific domain than any AI algorithm that has been trained on a (however large) data set of a particular domain. This is because the AI models can only make inferences based on a very low dimensional data set of the real world and are severely restricted (compared to human scientists) in the kind of modalities that they can process in their input data. These limitations are addressed by expert interpretation of the outcomes of these AI models. However, complications emerge when the inclusion of expert knowledge leads to the introduction of further uncertainties in the predictive capacities of the model. San Pedro (2020) talks about a trade-off between the inclusion of expert knowledge to reduce opacity which comes at the cost of uncertainties in the outcomes (San Pedro 2020, 16).

It can be argued that the experts themselves are epistemically inaccessible to the layperson, so are these experts no better than an epistemically opaque model? This is a conflation of the concept of opacity with inaccessibility. We always have a degree of inaccessibility between people, but testimony can still be a valid source of knowledge in epistemology. Without falling into the trap of philosophical skepticism, we can say that a human with sufficient expertise is able to understand and communicate domain-specific knowledge of their field to other educated non-experts. A complex and opaque AI model cannot do that.

However, some scholars claim that a lot of the complex and opaque elements of an AI model might be a feature rather than a bug of these computer programs. Alvarado (2021, 14) argues that if AI models were designed to be more understandable to humans, it would come at the cost of their predictive power; so much so that they would lose their competitive edge over other conventional methodologies and be unable to serve any useful function.

This point complicates our analysis because it undermines a fundamental assumption that explanation and understanding are intrinsic goals of science. But will these goals continue to remain central in science if they come at the cost of predictive capabilities? What is the nature of the trade-off between the epistemic goals of prediction and understanding? Some scholars, like Rudin (2019), would disagree with an inherent trade-off between various epistemic desiderata like choosing between explanation and predictions. However, they are in a minority. Moreover, different scholars present these trade-offs between explanation and prediction in their unique way, like the trade-off

17. Assuming that we were to ignore all ethical questions about the deployment of AI tools.

between accuracy and decomposability for Fleisher (2022), making it difficult to make claims about general trends outside of local contexts.

I will revisit these questions in the next chapter. Because before we can answer these questions, I will need to complete my review of the implications of opaque AI in science. In the next section, I will present a framework for understanding the various notions of opacity and define the appropriate usage that is necessary for advancing my argument.

4.8 A taxonomy of opacity in AI models

Facchini and Termine (2022) formulate a taxonomy to understand the different forms of opacity in AI models. This is helpful, as I have already seen how the technical concept of opacity is often conflated with other concepts. One common misconception is to conflate opacity with *any kind of epistemic inaccessibility* on the part of the agent towards a phenomena, be it a computer process, another human being, and so on.

Nevertheless, in a computer process, different forms of opacity arise due to various reasons. Facchini and Termine (2022) broadly distinguish between three of them. These are:

1. *Access opacity*, occurring due to the limitations of the human agents in accessing the epistemically relevant elements of a computer program that would allow for explanation, prediction, intervention, and so on. Access opacity is further subclassified into opacity of the training sample, the training engine, and the learned model.

2. *Link opacity* occurs when the AI model is employed to model a phenomenon wherein the model lacks the epistemically relevant elements that would allow it to explain, predict, and intervene on the target phenomenon. The link refers to the bridge between the model and the target phenomenon.

3. *Semantic opacity* concerns the storage and manipulation of information by AI models. As an example, AI models using deep learning algorithms to store information in the form of numerical functions and parameter values which do not have a straightforward semantic interpretation. Moreover, the processes by which the AI model manipulates data to make inferences might also be opaque to human agents.

Facchini and Termine (2022) further subclassify link opacity into 3 forms based on the fundamental notions of causation, mechanism, and law, giving rise to *causal opacity*, *mechanisms opacity*, and *laws opacity*. Causal opacity prevents AI models from

moving beyond observed data and learning the causal chains behind the phenomenon. Mechanisms opacity refers to the AI model's inability to hypothesize the existence of mechanisms behind the observed data in the form of more fundamental entities and processes.

The inability of AI models to identify laws is laws opacity. Even though AI models can easily train on large data sets to formulate "laws" like Charles's law and Boyle's law, there are other more fundamental laws in science. Laws in science can go beyond merely quantifying patterns of regularity in observed data. Laws also hypothesize the existence of theoretical entities and state the rules of interactions between them, using which we can deduce the observational generalizations. An example of this is how Kepler's three laws of planetary motion[18] can be derived from the more fundamental Newton's laws of motion.

4.9 Can AI-ML models be used in scientific practice?

Scholars of ML modelling in science also make claims about the necessary conditions that would qualify a model as a "scientific model." For instance, Fleisher says – "Scientific models are tools that we use to both understand and manipulate the world" (Fleisher 2022, 551). This lets us rephrase our original questions about epistemic trade-offs in a different form. Can a model that has:

1. High predictive capabilities but,

2. cannot provide explanatory understanding,

be accepted as a *scientific* model? Is prediction more valuable as an intrinsic epistemic goal than explanation or understanding? How do we choose between a predictive model with little explanatory value compared to an explanatory model that provides understanding but has little predictive value?

It is difficult to answer these questions because different domains of science demand different things from a model. Models in various domains like physics, biology, economics, engineering, and so on have their own unique epistemic demands and priorities. Moreover, as these ML models are adopted by different domains, the necessary requirements of what constitutes a scientific model for that domain can also change.[19]

It will be useful to test our intuitions with a thought experiment involving systems that we do not understand, but that nonetheless produce perfectly accurate predictions. Although no model or system can achieve perfect predictive accuracy, nevertheless, the

18. Kepler's laws were derived from discerning mathematical patterns from a large data set of astronomical data similar to how a lot of AI-ML models train themselves to identify patterns from large data sets.

19. Thanks to Varun Bhatta for emphasizing this point.

thought experiment will help us assess whether predictive accuracy alone is sufficient to qualify a model as a scientific model.

Symons and Alvarado (2019) present a similar thought experiment involving a "mechanical oracle". This raises for a non-pragmatic perspective science: is prediction by itself sufficient, even if it comes at the cost of theoretical understanding? Are theories and models an indispensable part of the practice of science, or are they elements of an ad hoc narrative that has been constructed by scientists to explain their practices?[20] Symons and Alvarado (2019) argue that:

> By stipulation, the oracle in our example has maximal predictive power. Therefore, in this case, there is nothing more for a pragmatist to do or to ask for. The inability for the pragmatist to explain the difference between the oracle and ordinary scientific practice is indicative of the weakness of pragmatism as a philosophical position. (Symons and Alvarado 2019, 44)

I will respond to this argument by noting that the oracle, despite its empirical success, lacks a notion of reliability and control, which can make it difficult for humans to trust the predictions. If it were to cease functioning, for whatever reason, humans would be left in the dark, as they would not be able to fix the oracle if they don't understand it. I will argue that this by itself can serve as sufficient motivation for a pragmatist to try and understand a perfectly predictive model (and if that was not possible, then to parallelly construct a theory-driven model) because of a lack of reliability in its future functioning.

4.10 Implications of AI opacity in scientific practice

Suppose reducing opacity is an epistemic virtue, and that explanation in the form of expert interpretation on the outputs of an AI model is a necessary condition for reducing opacity. Then scientists should construct AI models that are amenable to expert interpretation and subsequent explanation. This is supported by Rudin (2019), who advocates for constructing inherently explainable AI models rather than trying to devise ad hoc explanations for opaque models—explanations that are, in any case, severely limited and might not correspond to the real patterns that the AI has discerned in the data.

However, this idea of the possibility of reducing opacity by explanation is challenged by some philosophers who posit that constructing AI models in such a manner that they are more interpretable to humans will constrain these models and might come at the cost of their predictive power. Some of these philosophers go so far as to say

20. Thanks to GN for highlighting this point.

that these models will lose all their competitive advantage if they are made humanly comprehensible, rendering them useless (Alvarado 2021). However, I will again point out the difference between the *typical outputs* of an AI model versus the outputs of an AI model that can be accepted as *scientific predictions* or *discovery claims*. Even if a complex but opaque AI model vastly outperforms a more transparent model, the latter might still be preferred due to its ability in integrating its results into theory.

I am not arguing for the need to completely discard opaque AI models. There will be a degree of opacity in all AI models and many opaque AI models are being justifiably used in various domains of science (Duede, 2023). Rather, I am saying that experts will necessarily keep on trying to explain their AI models on their own terms, even when these AI models become increasingly complex in order to make accurate predictions. Even when an AI model is the best way to accurately model a system, experts will try to construct AI models that are less opaque and more open to expert interpretation in order to gain additional epistemic desiderata other than predictive capabilities.

By definition, the inner workings of an opaque AI model will resist any direct explanations by an expert. Nevertheless, scientists continue to use anthropocentric concepts to interpret the outputs of an AI model.[21] For example, scientists will attempt to demonstrate how the outputs of an AI model can also be inferred using more fundamental theoretical entities in an ad hoc manner (Duede 2023).[22] This prompts us to question the difference between:

1. Bottom-up and *data-driven elements* like numerical functions and parameters created by the AI model from a particular data set, and

2. the relatively stable and theoretically situated *anthropocentric elements* like general concepts, scientific explanations, and various other theoretical entities of science.

In the rest of this chapter, I will focus on one such anthropocentric element of scientific explanation. Although explanations are an intrinsic epistemic goal of science (Woodward and Ross 2021), an explanation also has various other instrumental goals. The authors in Srećković, Berber, and Filipović (2022) posit one such epistemic desiderata of the affordances that explanations provide scientists in making more and better predictions.

I want to critically review the idea of AI models being able to make predictions without explanation (Boge, Grünke, and Hillerbrand 2022). This will allow me to subsequently examine whether predictive AI models with high accuracy will make explanations obsolete (Srećković et al. 2022).

21. Note that philosophers and scientists alike highlight flaws in the practice of ad hoc theorization (Alvarado 2021; Rudin 2019).

22. I will revisit the notion of ad hoc explanations in chapter 6 and contrast it with the more epistemically neutral notion of post hoc.

The wider debate that this thesis is situated in concerns the challenge being posed by the adoption of AI models in science by critically examine two contrasting positions:

1. Do explanations, theoretical concepts, and other anthropocentric elements of science serve as mere metaphorical crutches for human scientists—constrained as they are by their limited cognitive faculties, or

2. are these elements indispensable tools that allow scientists to go beyond the immediate context of their predictions in order to have a broader understanding of the world, which is a necessary requirement for the progress of science? This view will support the position that explanations are a feature of scientific activity and will not be rendered obsolete by the predictive success of opaque AI models.

4.11 Explaining AI models

Although AI models are constructed by scientists, they themselves do not completely understand the inner workings on these models. This might threaten an anthropocentric way of conducting science where creativity, intuition, and insight are placed center stage in the process of prediction and discovery. It is hard for scientists to trace the outputs of these models to a particular input, or find connections between the outputs and theory. The problem compounds when we need to use these AI models outside of the domain of their creation. An AI model might be excellent in quantifying the quality of loan applications in the United States, but can we use the same model to sort applications in India? How can we explain a particular decision that was made by the model in accepting or rejecting individual loan applications? The answers to these questions are not clear, at least when we are dealing with typical AI models (Kleinberg and Mullainathan 2015).

The field of eXplainable Artificial Intelligence (XAI) has developed in response to the problem of opacity and the lack of interpretability of AI systems. XAI is an interdisciplinary field that aims to render AI systems less opaque and more understandable (Facchini and Termine 2022).

San Pedro (2020) argues against the idea that a process P is inherently opaque by noting that it is always possible to modify the process P to P', where P' is similar to P, but is also more transparent than P. San Pedro's formulation of opacity is such that we can actually utilize domain expertise and knowledge of accepted theory to completely eliminate opacity.[23] However, I should note that expert knowledge is (by definition) human-comprehensible. However, modifying AI models as to make them more amenable in incorporating inputs from expert knowledge can come at the cost

23. Please refer to the example of the simple pendulum by San Pedro in section 4.6.

of their predictive power. San Pedro (2020) solves this puzzle by asking researchers to strike a balance between the use of expert knowledge and the amount of uncertainty that the use of expertise introduces into the model.

I should note that any explanation that is associated with an AI model will be greatly diminished in its predictive power compared to the AI model itself. If this were not true and if the explanation could perfectly account for every prediction of the AI model, then we would not require the AI model in the first place! We could easily replace the AI model with the devised explanation. As Rudin (2019) herself puts it:

> Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (Rudin 2019, 207)

Nevertheless, one could always build models that are inherently explainable, which won't force us to use ad hoc and inaccurate explanations to understand and justify these AI models. On the subject of ad hoc theorization, Johnson and Lenhard (2024) note that ad hoc theorization is prevalent (and justified) in scientific practice beyond ML modelling:

> Analysis of recent and historical examples has led philosophers of science to a nuanced viewpoint that has partly revaluated ad hoc measures. Various studies have shown convincingly that ad hoc modifications form an important part of scientific practice that does not undermine confirmation. (Johnson and Lenhard 2024, 185)

Nevertheless, we need to be skeptical of what specific ad hoc practices in ML modelling are epistemically warranted, and which ones are a guise for psuedoscientific practice. Because these AI models cannot explain their own results, there's an onus on us to try to understand the reasons behind the predictive success of these AI models. AI models can help us identify subtle patterns in data that can aid us in improving our theoretical knowledge of the domain. Further research work should aim at reviewing the field of XAI and analyzing the concept of explanation that is central to the nomenclature of the field. This pursuit can further benefit from insights found in the philosophy of science literature that seeks to characterize opacity in contemporary AI models.[24]

To this end, the next chapter offers a review of a triad of the fundamental concepts of explanation, prediction, and discovery in philosophy of science. My focus will be in exploring the connections between the different elements of this triad.[25] I will argue

24. Thanks to Varun Bhatta for guiding me to this point.
25. Thanks to Varun Bhatta for emphasizing this point.

that philosophers of ML modelling in science and scholars in XAI have paid insufficient attention to the individual concepts of prediction and discovery in ML modelling along with their linkages to the concept of explanation. I will draw from standard frameworks in the philosophy of science to analyze how AI-ML scholars use these concepts.

CHAPTER 5

THE TRIAD OF EXPLANATION, PREDICTION, AND DISCOVERY

5.1 Explanation

5.1.1 What is an explanation?

A scientific explanation can be conceptualised as something that goes beyond a mere description of a phenomena.[1] An explanation attempts to understand a phenomena by invoking theoretical elements like unobservable entities, mechanisms, natural laws, and so on, which can help us connect the individual phenomena to theory.

Explanatory power seems to be a necessary requirement for any theory and it is hard to imagine a theory that has no explanatory value. Explanatory power is one criterion that can help philosophers demarcate between theories as opposed to other structures of organizing information which do not provide any overarching causes, effects, or connections.

5.1.2 Different philosophical accounts of explanation

More extensive study of the concept of scientific explanation was started by the introduction of the DN (Deductive-Nomological) model of explanation by Hempel (Woodward 2014). The DN model utilizes the concepts of *explanandum*, or what is to be explained, and the *explanans*, or that which is used in explaining the *explanandum*. To explain is to subsume the explanandum under the appropriate explanans. This is the reason why the model is called the deductive *nomological*[2] model, wherein an explanation

1. My understanding of the topic has been greatly shaped by the extensive works of Sundar Sarukkai, including his books and online courses.

2. The meaning of the word can be roughly translated to "law-like," which emphasizes the significance of natural laws in scientific explanations.

attempts to derive particular phenomena by way of inference general natural laws. At the heart of the DN model is the idea of *nomic expectability* (Woodward 2014). This is the intuitive acceptance of the existence of natural phenomena if they can be demonstrably derived from general natural laws. The foundational assumption involves accepting the metaphysical existence of these general laws.

There are various criteria that need to be satisfied for the explanation to be valid; for instance, although the explanandum has to be a logical consequence of the explanans, it should be so in a nontrivial way. The explanans should incorporate laws whose inclusion is necessary for the explanandum to be a consequence of the explanans. This guards against unfalsifiable explanations that can be formulated and derived in an ad hoc manner.[3] The DN model is complimented by the IS (Inductive–Statistical) model wherein the explanans can include statistical laws. These statistical laws can explain individual phenomena by demonstrating the likelihood of the occurrence of the phenomena (Woodward and Ross 2021).

In sharp contrast to its popularity, it's interesting to note that the DN model is no longer held in any regard among contemporary philosophers of science. This is due to its inability in adequately addressing its critics who highlight its overly rigid structure along with the fact that the model has struggled to explain situations where complete laws aren't known. Moreover, there is also the failure to address the nature of explanation in observational sciences where explicit causal relationships aren't always known (Woodward and Ross 2021).

Woodward (2014) and Woodward and Ross (2021) summarize the different accounts of explanations that differ on the basis of the epistemic virtue that they emphasize (among other things). These are:

1) *Causal* models of explanation: This set of approaches focuses on identifying the causal mechanisms that produce a phenomenon. Understanding these causal relationships is of central importance in providing a satisfactory explanation. The theory of evolution by natural selection as proposed by Charles Darwin is a prominent example of a causal model of explanation.
2) *Unification* models: These models identify the epistemic goodness of explanations that connect diverse phenomena under a single, unifying framework. Unifying explanations are often seen as more elegant and powerful than those focused on isolated phenomena. A popular example of this model would be Newton's unification of the seemingly disparate phenomena of terrestrial and celestial motion under his laws of motion and gravity.
3) *Pragmatic* theories: This perspective considers factors beyond pure logic and evidence that influence the judgement of explanations. Pragmatic theories of

3. There are important and widespread concerns about ad hoc explanations and hypothesis formulation, and I will give more attention to the "ad hoc" adjective in the section on prediction.

explanation diverge from conventional approaches by emphasizing the indispensable role of context and the psychological attributes of both the explainer and the audience. Additionally, it recognizes that a single phenomenon may have different explanations depending on the disciplinary lens through which it is examined. For instance, suppose a person jumps off of the top floor of a building. When asked to "explain" the problem, a physicist can only idealize the phenomena as a mechanical problem of a body falling with some mass m and some velocity v; whereas psychologist will study it from the (correct) perspective of human behavior and mental health.

Limitations of philosophical accounts of explanation

Each of the different accounts of explanations has its limitations; limitations which have themselves been well documented and categorized. I will mention just two, which are the problem of explanation asymmetries and explanation irrelevancies, by illustrating them with two popular examples as noted verbatim in Woodward & Ross (2021).

The problem of *explanation irrelevancies* notes how a derivation can satisfy the DN criteria and yet be a defective explanation because it contains irrelevant elements besides those that are associated with the causal features of the explanation. For instance, look at the following derivation:

Premise 1: All males who take birth control pills regularly fail to get pregnant.

Premise 2: John Jones is a male who has been taking birth control pills regularly.

Conclusion: John Jones fails to get pregnant.

Despite satisfying the DN criteria this is obviously not a valid explanation.[4]

The problem of *explanation asymmetry* highlights the many cases in which a derivation of an explanandum from a law using some initial conditions seems explanatory, but a "backward" derivation of the initial conditions themselves using the explanandum and the same law does not seem explanatory. This demonstrates the need to include causal elements that provide directionality to the structure of an explanation. For instance, one can derive the height of a flagpole by using:

1) The length of its shadow and

4. Note that there are ways in which the additional criterion on the explanans could be modified to make the DN model more strict with respect to a range of explanations. However, these additions seem to be made in an ad hoc manner (Woodward and Ross 2021).

2) the angle of the sun above the horizon.

Which is absurd because it is actually the height of the flagpole that is responsible for the length of the shadow, and not the other way round.

Other open questions in the philosophy of explanation concern discerning structures that might be common to all scientific explanations. However, some scholars hold that a scientific explanation is embedded in whatever specific disciplinary framework that is employed in studying a phenomenon.

5.1.3 Explanations, theory, and AI

The pragmatic accounts of explanation emphasize the significance of the cognizer and the context in which the explanation is framed. This leads us to question whether an AI cognizer be able to provide or receive an explanation? And if it does, what is the nature of explanations that are formulated by AI?

One possible answer are the structures of ML models. However, note that frameworks created by ML models can signal a break away from theoretical understanding of humans. Newton-Smith (2000) describes a thought experiment wherein an alien civilization could come to accept a set of scientific theories that are of a very different nature than our own. Newton-Smith is using this example to illustrate a popular concept of the underdetermination of theory by data. The argument considers the possibility of constructing a theory that is incommensurable with our current theories but which can nevertheless prove to have similar (if not more!) predictively capabilities. With the rising predictive success of ML modelling, there are concerns that the further development of AI can represent such a breakaway point for "human science".

It would be interesting to compare the "theories" made by an AI and those made by humans. But what would it mean for an AI to make its own theory? What is theory anyways? Does the expertise and technical know-how of an experimentalist also come under the domain of theory? It is hard to draw a distinction between theory and observation, a distinction that is further distorted by ideas of the theory-ladeness of observations.

In light of this, I think it is pragmatic to set aside the concept of theory and explanation for the time being, and analyze the concept of prediction. Before comparing AI theory and AI explanations, it is more fruitful to compare the predictions made by an ML with those that are made by a human scientist.[5] Although it is possible to easily compare the predictive capabilities of an ML model and a conventional theory-driven model; the more difficult question of the epistemic status of computer-generated

5. In this context, the predictions of a human scientist also include the predictions of a theory or a non-opaque model that is constructed by human scientists.

elements of an ML model remains unsolved.[6]

While analyzing Duede (2023)'s work, I will demonstrate that scientists are revising extant theory on the basis of the outputs of opaque ML models. This raises questions about the value of scientific theory that is modified on the basis of the outputs of ML modelling. If a scientist endorses[7] a particular theory on the basis of the outputs of ML models, are they merely using ad hoc accommodations which should undermine the epistemic status of the modified theory?

As the field of XAI[8] has boomed, a lot of work has been conducted by philosophers to investigate the nature of explanations in the field of AI. Therefore, I do not deem it necessary to provide a more comprehensive review of the concept of scientific explanation as it was understood before the popularity of ML modelling. What I will show in the next section is how the concept of explanation in ML modelling is fundamentally linked to the concept of prediction, and it is the concept of prediction that has lacked philosophical scrutiny.

5.2 Prediction

5.2.1 What is a prediction?

The concept of scientific predictions is widely debated in philosophy; including debates about the nature of predictions across different domains like predictions in physics versus predictions in biology. The concept of prediction is also associated with ideas of power, manipulation, and intervention over the natural world. This highlights the dynamic and multifaceted nature of the concept. Since ML models are primarily touted on their predictive capabilities; it will be helpful to review the history and various different notions of the concept of prediction.

I will posit that the concept of prediction in science cannot be divorced from that of explanation. This is because the justification for a prediction does not solely rest on empirical adequacy but also on its coherence with theory. Connecting each individual prediction to broader theory ends up justifying the prediction itself.[9]

In this vein, scientific predictions flow downward from *scientific theories* and *models* (Forster 2014). Thus, scientific predictions are inferences concerning possible observations (often about the future) that are justified on the basis of the postulates that lie behind the prediction. These postulates can be back-traced to the acceptance of

6. If these ML models continue to improve on their predictive capabilities, will that mean that the underlying (mathematical-logical) architecture of the ML model can also improve extant theory?

7. Please refer to Schickore (2022)'s work for an explicit review of endorsement as a "weak" form of justification. I will come back to these sets of ideas in chapter 8.

8. eXplainable Artificial Intelligence (XAI).

9. I will return to these sets of ideas in chapter 7 and 8.

scientific theories and models in a domain. In the context of ML modelling, a scientific prediction is a scientifically justified output of a model.

Another notion of prediction refers to a model's ability to forecast phenomena or results that have not yet been observed. A successful prediction demonstrates that the theory has explanatory and heuristic power. In this vein, the test of theory becomes the correctness of its predictions. A common theme that is explored among scholars is the contrast between *novel predictions* and *accomodations*. *Novel predictions* are about phenomena that were unknown at the time the theory was formulated and are a strong indicator of a theory's validity. On the other hand, *accommodation* involves the practice of explaining already known phenomena by modifying the under-construction-theory.

Prediction is also linked to the idea of confirmation, where confirmation refers to the fundamental idea of how evidence supports (or *confirms*) a particular hypothesis. Confirmation can be understood in an objective or a subjective manner (Forster 2014). From an objective view, confirmation is simply a logical relation between a hypothesis (along with the background assumptions of the extant theory) and evidence. However, the degree of prior belief in a hypothesis before the confirmation test can vary with different cognizers. Therefore the "degree of confirmation" becomes a variable and introduces a notion of subjectivity in the conceptual analysis of confirmation. This subjective Bayesian perspective on prediction reduces the act of prediction into the determination of a numerical value that will represent the probability of the occurrence of an event.

Although Foster (2014) advocates for formulating better philosophical accounts of predictions that will allow us to move beyond merely evaluating our subjective degree of belief; nevertheless, a study of prediction will also have to be sensitive to the above-mentioned features of confirmation that will be inherited by the concept of prediction.

Predictions can also be distinguished on the basis of their probabilistic or deductive nature (Forster 2014). As I have already mentioned, prediction can be conceptualised as a logical relation between theories, hypothesis, and evidence/observations. However, in actual scientific practice, predictions are rarely deductive and are mostly probabilistic. In order to quantify the strength of a prediction, a Bayesian framework is employed to assess the assigned probability of the actual occurrence of the predicted event by theory. In contrast to this, another form of prediction is rule-based prediction where predictions are made on the basis of established rules; regardless of how a Bayesian probability framework can be applied to this set of rules.

While Forster (2014)'s review on prediction describes the various logical relations between predictions and other concepts like confirmation, theory, evidence, and so on; the central theme in the more comprehensive review by Barnes (2022) is that of differentiating between prediction versus accommodation and how the former is

assigned a higher epistemic status than the latter. This common epistemic intuition where we assign a higher epistemic status to prediction as opposed to accommodation is called *predictivism*. More precisely: predictivism entails that evidence confirms theory more strongly when predicted than when accommodated.

5.2.2 The notion of temporality in predictions and the epistemic nature of ad hoc hypothesis

This brings me to another fundamental notion associated with prediction: the notion of temporality and forecasting associated with the formulation of a prediction claim. Here, it will be useful to draw a finer differentiation between *post hoc* hypothesis and *ad hoc* hypothesis. *Post hoc* roughly translates to "after this", and refers to something being applied retroactively. *Ad hoc*, on the other hand, means "for this purpose" and describes something created or done to address a specific issue or situation, and has the connotations of being an improvised and temporary solution. While the phrase post hoc deals with retroactive applications, ad hoc focuses on case-specific responses. Similar to how accommodation has a lower epistemic status than prediction, the construction of ad hoc hypothesis (including ad hoc explanations) by a particular scientist is met with skepticism by other scientists.

But what does an ad hoc hypothesis mean? And how are ad hoc hypothesis different from hypothesis that are imbued with a stronger predictive element? The definition of the phrase "ad hoc" suggests that rather than the nature of the hypothesis itself, what makes a hypothesis an ad hoc hypothesis is the *intention* behind the construction of the hypothesis. Many scientists construct ad hoc hypothesis to explain away a specific anomaly without having to disband their entire theory.

However, Popper thought that these ad hoc hypothesis are suspect not because of the intentions of the scientist, but because of the nature of the hypothesis itself. For Popper, ad hoc hypothesis could not be tested independently of the phenomena that they were "saving". This is because there are no other testable consequences of accepting a particular ad hoc hypothesis (Barnes 2022).

The criteria of lacking avenues of independent confirmation is helpful in demarcating ad hoc hypothesis, and Barnes (2022) refers to this meaning of ad hoc as ad hoc_1. However, there are other criterion that can be also used to understand and demarcate ad hoc hypothesis. As Barnes (2022) notes: A hypothesis introduced to accommodate a set of data could be qualified as an ad hoc hypothesis simply because it was an unconfirmed hypothesis (ad hoc_2); or if it failed to cohere with the basic commitments of the research programme in which it is proposed (ad hoc_3).

For this thesis, my usage of ad hoc is that of ad hoc_3. I do not want to analyze the motivations behind the formulation of a proposed hypothesis. I simply want to

point out that ad hoc hypothesis might not be within the foundational framework of a research programme (or paradigm), and if this is true, then the hypothesis might simply be false.

Ad hoc hypothesis or explanations are constructed to save a theory from being falsified. This often leads to difficulty in testing these hypothesis. However, post hoc hypothesis are not to be percieved as be being imbued with malicious intents of saving the phenomena. For this reason, temporality alone should not be a criterion for demarcating ad hoc and non ad hoc hypothesis. Rather than temporality, what raises suspicion for an ad hoc hypothesis is the inclusion of arbitrary concepts that lack independent justification for their existence other than simply accommodating an anomaly.[10]

However, scholars lack consensus on whether the phrase ad hoc is a useful concept in describing scientific practice. This is because the scientific community's judgment about whether a hypothesis is ad hoc can change with circumstances:

> Given this revisability, and the aesthetic dimension of theory evaluation (which leaves assessment to some degree 'in the eye of the beholder') there may be no particular point to embracing a theory of ad hocness, if by the term 'ad hoc' we mean "illegitimately proposed". (Barnes 2022, citing Hunt 2012)

Similar to ad hoc hypothesis, ad hoc explanations are also a cause for concern in science including our current focus which is the field of ML modelling. But yet again, the distinction between the phrases of ad hoc versus post hoc will be helpful. While ad hoc explanations could very well be flawed, Fleisher (2022) attempts to defend ML modelling against the *rationalization objection* of post hoc explanations in ML modelling. Critics posit that any explanation that is devised for an opaque ML model is a mere rationalization and can never be a genuine explanation. This is because the opacity of a model will not allow experts to assess whether the devised explanation is faithful to the inner workings of the original opaque model. As I have already mentioned, this problem has recieved widespread attention by scientists and philosophers alike (Rudin 2019; Sullivan 2023).

However, I will propose that Fleisher (2022)'s formulation of the problem could also be viewed as a crisis of predictions rather than of explanations. Parallel to Fleisher's rationalization objection, it is possible to construct an *accomodation objection* against ML models wherein I argue that ML models are constructed by over-fitting on training data and their outputs are not faithful predictions. The argument can be extended to make the radical claim that ML modelling practice is one that produces ad hoc accomodations

10. Similar concerns hold true for ad hoc explanations.

and these models cannot be said to make scientific predictions. I will revisit this idea in more detail in chapter 7 by analyzing a case study.

The notion of novelty in prediction

Another important notion associated with the concept of prediction is that of novelty. Novel facts and observations provide a higher degree of confirmation to a hypothesis, and the reason why a prediction enjoys a higher epistemic status than accommodation is because predictions make claims about novel, as yet unobserved facts. This motivates the need to review the concept "novelty" similar to our review of temporality in the previous section.

Barnes (2022, citing Zahar) defines a fact to be novel if it was not part of the problem space that originally led to the construction of a hypothesis. Another conception suggests that evidence is qualified as novel if the evidence ends up supporting a theory even if the theory was not constructed to fit (or accommodate) the evidence.

The two different conceptions of novelty are called *problem-novelty* and *use-novelty*. *Problem-novelty* refers to a fact being novel if it was not part of the original problem that guided the construction of a theory. In contrast, use-novelty states that a fact is novel if the theory that confirms the fact was not specifically designed to accommodate it, regardless of whether the fact was part of the initial problem situation. Use-novelty focuses on whether a theory predicts a fact independently, whereas problem-novelty concerns whether the fact was among those originally considered during theory formation. This terminology offers a useful framework to address a debate concerning the reasons behind the predictive success of ML models. Are ML models making *novel predictions*, and if they are, are the outputs of an ML model problem-novel, or use-novel? The reason why the debate has not been settled is because it is hard to know what theory should be invoked in understanding the inner workings of an ML model.

I will now try and show which account of predictivism is most suitable for our analysis of ML modelling. This will depend on which account of predictivism in contemporary philosophy of science best captures the debates in ML modelling.

5.2.3 Various accounts of predictivism

The various contemporary accounts of predictivism respond to a famous thought experiment by Maher (as cited in Barnes 2022). In order to study our intuitions regarding prediction and accommodation, Maher proposes a thought experiment with a simple setup wherein a coin is tossed many times and each of the outcome is noted.

Suppose a theory T claims that a biased coin will always land on heads. A predictor establishes T *before observing any evidence*, and after witnessing 99 heads in a row, the

100^{th} toss further confirms T. In contrast, an accommodator *first observes* the 99 heads and then formulates T to fit that data, subsequently using T to predict the 100^{th} toss.[11]

While both approaches lead to the same claim, the predictor is percieved to be *epistemically stronger*.[12] For the predictior's case, T is continuously tested and confirmed throughout the process, whereas for the accommodator, T is only explicitly tested on the final toss. The key difference lies in *when*[13] T is established by an agent relative to the agent's observation of the evidence. This alters our intuition regarding the confirmation of two theories with the same predictions by the same set of observational data if one is predicted and the other is accommodated (Barnes 2022).

Predictivism will posit that the predictor has a stronger epistemic claim than the accommodator. But why is that? One explanation that warrants skepticism towards the accommodator highlights the possibility of the accommodator fudging their hypothesis in light of new evidence. The fudged hypothesis (or explanation) might not cohere with theory and would end up being an ad hoc hypothesis (Barnes 2022).

Another account of predictivism is given by Lange (as cited in Barnes 2022) wherein they claim that theories arising as a result of an accommodation process are not strongly supported by confirmation.

Lange argues that predictivism is appealing if we assume that all the previous 99 tosses are independent of each other. However, this might not be true. If a predictor can successfully predict the 99 outcomes, it provides evidence that they have uncovered an underlying pattern that was not obvious beforehand. All the 99 coin tosses were not independent of each other, and what is important is to discover the underlying cause. For Lange, whether the underlying cause is determined by a predictor or an accommodator is besides the point. This is because the issue is not merely whether the outcomes are predicted or accommodated, but whether the hypothesis is arbitrary or not (Barnes 2022).

Another concept that is used to understand our intuition about predictivism is that of a *severe test* that was introduced by Mayo (as cited in Barnes 2022). Mayo posits that the confirmation of a hypothesis depends on whether it has passed a *severe test* where a severe test is one that a hypothesis is unlikely to pass if it is false. Predictivism draws its intuitive appeal when novel predictions confirm severe tests, especially when these predictions are unexpected.

We also have Worral (as cited in Barnes 2022) invoking Duhem's account of con-

11. The setup has some resemble to the training of an ML model wherein a model learns from a training data set (the first 99 outcomes of the coin toss) and is tested on a sample data set (the outcome of the 100^{th} coin toss), supposedly outside the training data set.

12. An epistemic agent is *epistemically stronger* if they are perceived to have more justification for a particular claim; this can involve the agent being in possession of a superior epistemic process of generating said claim.

13. Yet again, this highlights the significance of the notion of temporality in understanding the concept of prediction.

ceptualizing scientific theory as a core set of claims combined with a set of auxiliary free parameters. When new evidence is used to modify a theory, it is mostly targeting these auxilary free parameters and does not challenge the core set of claims. Worrall's main argument is that the evidential weight of a prediction depends on the logical relationship between theory and evidence, rather than simply on whether the evidence was temporally novel. Although the evidence that is merely used to modify the auxilary parameters certainly affords a level of confirmation, the evidence that has implications for the core theory provides a stronger confirmation for the hypothesis.

Here, Worrall (as cited in Barnes 2022) supports a weak form of predictivism wherein the notion of temporality that distinguishes prediction and accommodation is not as important as whether the evidence directly impacts core theory or the auxiliary parameters. Weak predictivism does not claim that predictions are inherently superior to accommodation; rather, weak predictivism states that scientists should focus on trying to uncover the causal factors behind the epistemic success of a prediction. For instance, to determine if predictions usually end up capturing something in core theory rather than merely modifying with the auxiliary free parameters.

Worral's argument is a good segue to generalize a set of arguments for predictivism. An argument for predictivism that invokes the quality of the predicted theory being true can be framed in the following manner:

> A common argument for predictivism is that we should avoid inferring that a theory T is true on the basis of evidence E that it is built to fit because we can explain why T entails E by simply noting how T was built—but if T was not built to fit E then only the truth of T can explain the fact that T fits E. (Barnes 2022)

As Barnes notes, the interesting explanandum is not the fact that T entails E, but that the theorist was able to construct T in such a manner that T correctly entailed E. This suggests that the theorist[14] may have captured some underlying pattern or principle that guided them towards the true theory.

Another perspective on predictivism is provided by Hitchcock and Sober (as cited in Barnes 2022) wherein they start off by describing the trade-off between the goodness of fit versus simplicity in a mathematical curve-fitting problem. Beyond a point, goodness of fit with data is often sacrificed so as not to overfit the curve on the given data and use as few free parameters as possible. Hitchcock and Sober use this framework to advocate for a form of weak predictivism by incorporating the concern that an accommodator is more prone to overfit a theory. If a theorist accommodates data in building a theory, there is a risk of overfitting, making the theory less reliable. However, if a theory is developed independently of data and still manages to successfully predict the data,

14. Note the parallels between the successful theorist and a successful ML model.

then we have reason to believe that the theory was not overfitted and should enjoy stronger epistemic support compared to the theory constructed by an accommodator.

Hitchcock and Sober suggest a solution to the issue at hand. They formulate a process by which we can compare the theories made by a predictor versus an accommodator. One way to do this is by determining the criterion of epistemic strength (like the Alkaike criterion[15]) for both theories. If it turns out that the criterion is the same for both theories, then the history of how these theories were constructed should not matter. However, in practical settings, it might be difficult to actually determine and evaluate the criterion of epistemic strength for a particular theory. In these cases, the fact that one theory was predicted while the other was accommodated can serve as a proxy for their epistemic criteria.

In the review article, Barnes (2022) themselves introduce the weaker concept of an *endorsement* (as opposed to a confirmation) for theory evaluation. What matters is not the evaluative history of how a theory was confirmed, but rather the basis on which scientists endorsed a particular theory in the past.

Suppose that a theory entails some hypothesis, and a novel prediction confirms this hypothesis. Subsequently, the novel prediction now constitutes new evidence for the confirmation of the theory. Now, if a scientist had endorsed the theory without appealing to the new evidence, then the new evidence adds epistemic value for the scientists' subjective assessment for the original theory (Barnes 2022). Using this idea, Barnes advocates for a more useful distinction between *virtuous endorses* and *unvirtuous endorses*, rather than the distinction between predictors versus accommodators. A *virtuous* endorser assigns probabilities based on evidence and background beliefs, while an unvirtuous endorser is heavily influenced by social pressure, personal biases, career incentives and so on. I think this is helpful because it allows us to articulate the skepticism surrounding accommodators by explicitly noting that they are being potentially unvirtuous endorses.

However, it should be noted that not all scholars accept preditivism. *Anti-predictivism* challenges the idea that predictions always provide stronger confirmation than accommodations. Some scholars state that what matters is not whether a theory predicts data but whether it has independent reasons for supporting it. Critics of predictivism also cite historical data which suggests that predictivism exaggerates the

15. The Akaike Information Criterion (AIC) is a statistical tool used to compare models by balancing goodness of fit and simplicity. It is defined as:

$$AIC = -2 \log L + 2k$$

where L represents the likelihood of the model given the data, and k is the number of adjustable parameters in the model. A lower AIC score indicates a model that achieves a good fit without excessive complexity. Hitchcock and Sober identify AIC as a useful criterion by which the epistemic status of theories—whether predicted or accommodated—can be compared.

importance of predictive success in theory acceptance (Barnes 2022). Some go so far as to posit that accommodations can have a higher epistemic status than predictions because they might reduce the risk of data manipulation or fabrication!

5.2.4 Is the nature of a prediction defined by the nature of its premises?

I will offer a different perspective in understanding predictions by contrasting them with mere conclusions. In philosophy, a conclusion is a proposition that is logically inferred from a set of premises. When these premises are the part of a larger scientific theory or the assumptions behind a smaller modelling practice, then the conclusion can be called a *scientific* conclusion. Moreover, if these conclusions have a notion of temporality that extends to the future, or if they make claims that go beyond the available observational evidence, then these scientific conclusions can be qualified as *scientific predictions*.

If we were to take this definition to mean that only propositions that can be back-traced to foundational assumptions should be qualified as predictions in science, then most outcomes of opaque AI models would not be called predictions.[16] However, scientists seem to accept that AI models can make scientific predictions. This is evident from the fact that the Nobel Prize in Chemistry for the year 2024 was awarded to the team behind AlphaFold. A Nature report on the same states that the Nobel prize has been awarded to this team for the "predictive abilities" (Callaway 2024) of AlphaFold. This complicates the appropriate usage of the term "scientific prediction" in the context of AI models.

Will the term scientific prediction be expanded to include all proposed claims about scientific observables? The term "scientific prediction" is used to refer to things like forecasting the results of experiments, claims about direct scientific observations, or the results of a modelling exercise that makes claims about the observable world. However, can the outputs of an opaque ML model[17] that makes claims about the observable world be called a prediction? If yes, how are these predictions different from predictions as they are understood in conventional philosophy of science?

Before offering my response to these questions, I will provide further context by reviewing arguments from a recent book by Johnson and Lenhard (2024) which traces the history of the concept of prediction. According to the authors, one important factor that has influenced the evolution of the concept of prediction is the adoption of novel technologies by scientists.

16. Owing to their epistemic opacity.

17. As we have already seen, an ML model can make claims about observations. However, the opaque nature of ML models can make it difficult for scientists to infer semantic interpretations of the inner workings of the model. This poses problems for providing epistemic justification for the claims.

5.2.5 A new culture of prediction?

Johnson and Lenhard (2024) discuss the epistemic status of prediction in conventional philosophical accounts by noting that:

> In many philosophical approaches, prediction counts as a sort of technical achievement, whereas explanation is seen as the more virtuous epistemological goal. And we have seen that working with adjustable parameters is oriented toward prediction, partly to the detriment of explanation. (Johnson and Lenhard 2024, 186)

Philosophical accounts of prediction have lagged behind technological advancements, especially after the widespread adoption of computers in scientific practice. This gap has widened after the development of opaque ML models that are surpassing the predictive capabilities of conventional computer models.

Johnson and Lenhard (2024) offer a broad account of 4 different *cultures of predictions* in the history of science and technology. A *culture of prediction* refers to the complimentary set of practices that surrounds the process of making predictions that go beyond the available mathematical tools and ideas. A culture of prediction also includes the interplay between mathematics, epistemology, technology, and social organization of a particular historical period. Although most instances of predictions would exist as a hybrid between 2 or more categories, nevertheless, these categories can be helpful in developing a vocabulary that allows us to differentiate the plurality of ways in which the concept prediction has been understood in history. The 4 different cultures of predictions are:

1) The *rational* culture: Which is grounded in the assumption that natural phenomena should be studied using mathematical laws. Predictions are inferences drawn from mathematical postulates, and phenomena are studied using mathematical analysis and equations.
2) The *empirical* culture: Which emphasizes measurements and experimental data to make predictions about phenomena.
3) The *iterative-numerical* culture: This culture emerged with the advent of computers which were used to approximate mathematical equations of models using iterative algorithms. Scientists now had the ability to make predictions without needing exact analytical solutions for their equations.
4) The *exploratory-iterative* culture: Developments in technology allowed easy access to computational resources. This computing power was used to iteratively explore the parameter space of computer models. In this culture, the computer models themselves became subjects of the predictive practice.

Johnson and Lenhard (2024) trace the historical evolution of the development of data-driven algorithms. They note how easy access to computational resources allowed scientists to employ adjustable parameters to extend their theoretical computer models. Theoretical models incorporated with adjustable parameters are highly versatile and can greatly extend the applicability of a theory to practical problems. The architecture behind data-driven, opaque ML models pushes this to an extreme limit. These models minimize the relevance of the theoretical components in a modelling exercise along with undermining the role of domain expertise. For example, the choice of the mathematical-logical architecture for an ML model (for instance, the choice between a deep neural network, a decision tree, or a random forest) becomes more important than theoretical considerations from domain expertise.

In the concluding chapters, the authors hint at the emergence of a fifth culture of *pure predictions* surrounding AI-ML modelling in science (Johnson and Lenhard 2024, 197). However, they note that philosophers need to develop better conceptual frameworks to understand the meaning of pure predictions. The authors themselves state that these predictions are pure because they are divorced from any explanatory value or theoretical understanding and are only seen as a technological unit that can be applied to practical problems.

A book review of Johnson and Lenhard (2024)'s work also echoes this possibility:

> . . . there is justification to expect that a new culture of prediction may indeed emerge given the novel tools of machine learning. However, as they [the authors] note, we must wonder whether such a culture would prioritize prediction over other goals of science, like explanation. (Boesch 2024)

5.2.6 Predictions in ML modelling in science

There is widespread acceptance that ML models are very successful at making predictions (Beisbart and Räz 2022, 2). A ML model is said to have high predictive power if it has the capabilities of making accurate predictions outside of its training and testing data set. The emphasis on the novel predictive capabilities of ML modelling is consistent with the philosophical dictum placing greater epistemic weight on predictions as opposed to accommodation (Barnes 2022). Given that we do not have epistemic access to the internal workings of an opaque ML model, an ML model is accepted solely on the virtue of its predictive capabilities.

One could argue that there is no significant difference between the predictions of ML models versus the predictions of theoretical models in science. This is because both types of models make predictions in a form that is understandable and testable by humans. Duede (2023) implicitly accepts this idea where he invokes the context distinction in the philosophy of discovery to posit that ML models can justifiably be

used in the context of discovery in formulating novel hypothesis.[18] By invoking the context distionction, the argument states that the history of how the hypothesis was formulated is irrelevant. What matters is how the hypothesis is tested, and if it happens to be true.

At this point, it is helpful to highlight that the broader debate that we are situated in concerns the status of opaque ML models in science. There is a difference between merely accepting the outputs of the model, versus accepting the models themselves. Suppose the outputs of an ML model are judged to be valid by comparing with emperical data. However, the model itself might have other properties that will disqualify it from being accepted as a scientific model. One such requirement could be about providing explanatory value to the expert.

For instance, let us provisionally accept Forster (2014)'s definition of scientific predictions as the observable consequences of accepting scientific theories or models. Here, the acceptance of the outputs of ML models as scientific predictions hinges on whether we want to accept the ML models themselves as scientific models. This is a challenging question, because it is hard to determine if we can classify an opaque ML model which is not amenable to a semantic interpretation by a domain expert as a scientific model. However, if the outputs of these ML models end up making reliable and accurate predictions for a long time, will this epistemic capability automatically qualify them as scientific models regardless of their opacity?

Future work can attempt to provide a better understanding of the possibility of divorcing predictions from other epistemic goals like explanations, understanding, and so on. In this regard, it will be interesting to construct thought experiments and analyze some case studies of a highly predictive model that provides no understanding. This can then be substantiated by considering some extreme cases of opaque, black-box models. In the next chapter, I will provide a preliminary meta-review of the epistemic trade-offs in the use of ML modelling in science to understand how scientists are prioritizing predictions with respect to the other epistemic goals in science. However, before we move to a more indepth analysis of prediction in ML modelling, I will complete the triad by offering a review of the concept of discovery in science and its linkages to prediction and explanation.

5.3 Discovery

5.3.1 What is a scientific discovery?

Scientific discovery is used to refer to both the process and the results of successful scientific inquiry (Schickore 2022). The act of discovery is different from simple, indi-

18. I will revisit the context distinction in more detail in the subsequent section on discovery.

vidual observations of phenomena. A discovery can be an observational generalization, or any such interesting patterns in natural phenomena, although a natural phenomena itself can also be accepted as a discovery in some contexts. Discovery is necessarily novel, and discovery can be both observational, as well as theoretical.

A central theme in scholarly works on the philosophy of discovery concerns the existence and the nature of a *method* or *logic of discovery*. Should discovery be understood as an imaginative act of insight that is impenetrable to rational analysis? Or is there a particular method by which scientists can reliably produce discoveries?

Other questions in the philosophy of discovery concern the nature of discoverable entities, as well as the nature of the cognitive agents that can be said to make discoveries. On the latter issue, there is no consensus on whether an AI can be credited with a scientific discovery. Framing the question in this particular format also brings out the social nature of discovery wherein a scientist is only credited to have made a discovery when the scientist's result is recognized as being novel and valuable by a community of scientists.

An important facet of the study of discovery concerns the justification for a discovery claim. Obviously, a scientists cannot claim that they make a discovery without providing any justification for the same. But what counts as justification for a discovery claim? Moreover, can the process behind the justification of a scientific claim be seperated from the process that led to the discovery of the claim? Or as Nickles puts it:

> There must be some degree of coupling between the modes of generating theories and criteria of epistemic appraisal. (Nickles 2000, 91)

Another perspective on discovery is provided by Kuhn (Schickore 2022). For Kuhn, a discovery is associated with modifications to the current paradigm of a field. Because a discovery involves novelty and might be inspired from anamolous phenomena, Kuhn posits that discoveries can even trigger a paradigm shift. In this perspective, discoveries (or at the very least, transformative discoveries) cannot be made in normal science as the anamolous property of a discovery necessitates modifications to the paradigm in order to normalize the discovery. Citing the "discovery" of oxygen over the period of 1774-1777, Kuhn acknowledges the difficulty in identifying a single point of discovery and argues that the reason behind the extended discovery process is due to the modifications in a paradigm that can accommodate the novel results.[19]

At this point, I should emphasize that contrary to common sensical notions, a discovery does not happen at a particular moment in time. A discovery is not an event, but a process, which can extend for decades after the initial formulation of the

19. Note that this only happens in fields with an extant paradigm. In pre-paradigm science, discoveries happen with the simultaneous exploration of a new phenomena and the articulation of a tentative hypotheses.

discovery claim (Schickore 2022). This extended process involves the wider community of scientists scrutinizing the discovery claim. A discovery claim is accepted as such only when the community of scientists find a consensus on the fact that a knowledge claim has proven to be provisionally stable. As Arabatzis notes:

> . . . there is more to discovery than a eureka moment. Discovery comprises processes of articulating, developing, and assessing the creative thought, as well as the scientific community's adjudication of what does, and does not, count as 'discovery'. (Arabatzis 1996, as cited in Schickore 2022)

This adds an element of intersubjectivity and brings out the communal nature of scientific knowledge. Therefore, an analysis of the concept of discovery can benefit from sociological perspectives. As Schickore says:

> Sociological theories acknowledge that discovery is a collective achievement and the outcome of a process of negotiation through which "discovery stories" are constructed and certain knowledge claims are granted discovery status. (Schickore 2022)

5.3.2 Logic of discovery

The enlightenment period raised optimism about the fruits of systematic inquiry. There was a popular idea about the existence of a general method that could reliably produce new results and transform human life and society (Nickles 2000; Schickore 2022). Today, AI and its predictive capabilities have revived interest in the idea of a *logic of discovery*. A *logic of discovery* refers to the idea that there exists a logic, or a systemic and rational method, of reliably generating new scientific hypotheses and theories. In order to make a discovery, a scientist simply needs to learn the scientific method and apply it onto their problem.

This idea aligns with the Baconian vision of a universal scientific method (Nickles 2000). For Bacon, the scientific method was understood as a method of discovery that would lead to justified claims about the natural world. His grand vision for the future of science involved the production and accumulation of novel knowledge.[20] Belief in a particular piece of generated knowledge is warranted by the very fact that it is the output of a reliable process of generating knowledge.

However, the Baconian view was undermined by the fact that even the supposedly pure and rational methods of discovery often produced knowledge that was limited and fallible. This led to the development of the canonical HD (Hypothetico-Deductive)

20. This perspective on discovery blurs the contemporary distinction between the context of discovery and the context of justification; an idea that I will revisit in the later sections.

model of understanding science. For the HD model, it was the hypothesis (rather than the method) that became the fundamental element in inquiry. The standard HD model describes scientific inquiry as a process of examining provisionally accepted propositions and trying to prove or disprove them. The noteworthy difference compared to the Baconian view was that the process by which the hypothesis was *formulated* was irrelevant.[21] What mattered most was how the hypothesis were *justified* (Nickles 2000).[22]

To better understand this distinction, Laudan (1981) introduces a broad distinction between two different paradigms of *generativism* and *consequentialism* in his seminal paper titled 'Why was the Logic of Discovery Abandoned?'. The emphasis given to the process of generating scientific knowledge (similar to the Baconian view) can be generalized and is refered to as *generativism*. In contrast, the HD method is a model under the paradigm of *consequentialism*, wherein a claim is supported when the claim's consequences are confirmed by evidence. As Laudan himself puts it, the paradigm of consequentialism justified a discovery claim in the following manner:

> If an appropriately selected range of consequences proved to be true, this was thought to provide an epistemic justification for asserting the truth of the theory. (Laudan 1981, 184)

On the other hand, the paradigm of generativism:

> . . . believed that theories could be established only by showing that they followed logically (using certain allegedly truth-preserving algorithms) from statements which were directly gleaned from observation. (Laudan 1981, 184)

The more traditional scientific methodologies were generativist, while most modern approaches to science align with consequentialism (Nickles 2000).

The context distinction

In 1931, the positivist philosopher Reichenbach introduced the popular *context distinction* in philosophy of science. This was the distinction between the *context of discovery* and the *context of justification* of a discovery process. The *context distinction* marks the distinction between the generation of a new idea or hypothesis and the subsequent

21. Although the concept of discovery is closely associated with the concepts of rationality and logic, discovery also involves notions of creativity and insight. Some philosophers claim that the discovery process imagined as the moment of insight in a scientist's mind is inaccessible to philosphical scrutiny.

22. There are various limitations to the HD model. Chief among them is the concern that the model gives no insight into how amateur scientists learn the process of making discoveries of their own. Just learning the skills of *justifying* discoveries is not enough, a scientist also needs to know how to *make* discovery claims.

justification of it.[23] Although scholars were obviously aware of this distinction before, what marks its significance in the 20th century were the associated value judgements for elements in the context of discovery.

The logic of discovery came under attack by the positivists, some of whom went so far as to say that the topic of discovery should be expelled from epistemology and should rather be studied by psychologists (Nickles 2000). Popper famously rejected the generativist paradigm in his book 'The logic of scientific discovery.'

Despite the positivists, some philosophers continue to study discovery. Data gathered from historical analysis showed that (at least in some cases) discovery is not catalyzed by a sudden inspiration but is actually a structured, intellectual and experimental process. These scholars believed that the concept of discovery was central to epistemology as it was the discovery process itself that was at the frontier of scientific inquiry (Nickles 2000).

Although scholars were unable to construct a single monolithic logic of discovery, nevertheless, discovery could not be reduced to a mysterious, imaginative, and seemingly random process. After the popularity of the context distinction, arguments in favour of the existence of a logic of discovery took two forms. One approach equates the logic of discovery with abductive reasoning, while the other conceptualizes it as a problem solving algorithm employing heuristic elements.

Logic of discovery and AI

In the phrase "logic of discovery", *logic* is understood as a set of rules, a method, or a structured account of reasoning involved in making reliable discoveries and advancing knowledge. Schickore (2022) presents two broad approaches in understanding the logic of discovery. One way is to understand the logic of discovery as a systematic account of the reasoning processes involved in knowledge generation. Another way is to conceptualize the discovery process as a problem-solving algorithm employing heuristic elements.

Some scholars wanted to formulate a rational account of the process of discovery that could effectively be used across different domains; these scholars wanted to minimize the perception of chaos and the notion of luck surrounding the concept of discovery (Nickles 2000). An account of this nature would involve going beyond a context-specific logic, interlaced with emperical and theoretical content, and to formualte a content-free logic of discovery.

One such attempt was made by Newell and Simon where they constructed a so

23. Whewell described the process of discovery as being split into three parts: these are the "happy thought", the articulation and development of that thought, and the testing or verification of it (Schickore 2022). The part of the process concerning the development of the thought was described using various accounts of pragmatic logic.

called "general problem solver". This method aimed to only use some general and content-free set of rules that could be iterated in order to construct knowledge-based and case-based "expert systems" in AI (Nickles 2000, 88). Although their system was limited in accomodating a diverse domain of problems, they conceptualised the concept of discovery in an interestingly novel manner.

Newell and Simon conceptualized discovery as a form of a general problem solving exercise. A discovery would involve searching through possible solutions in a problem domain using an algorithm that was guided by heuristic elements. Although a heuristic procedure is fallible and at best provisionally acceptable, nevertheless, it is an efficient set of rules that can narrow down large spaces of search. This allowed the inclusion of heuristics in the logic of discovery, which although being false themselves, did not necessarily undermine the rationality of the logic.[24] This essentially reduced the discovery process to a "problem-solving" algorithm searching through possible solutions (Nickles 2000).

Another account conceptualizes the process of discovery as a problem-solving algorithm comprising a method of blind variance and selective retention. Popper posits that the process of discovery is similar to a form of darwinian selection wherein a problem is solved by offering iterations of possible solutions. With each iteration, the stronger candidate solutions are retained and subsequently varied in the next iteration (Nickles 2000).

As a closing argument, Nickles (2000)'s posits that there is no single universal logic of discovery in science. There exist particular methods that aid in discovery, however, these methods are practical and context-specific. These methods evolve with time and constitute heuristic elements and appraisals. Moreoever, it is hard to differentiate between the logic that actually led to the formulation of a discovery claim versus the post hoc[25] logic of rational reconstruction of a discovery claim. If a logic was only discernable *after* the formulation of the discovery claim, is it acceptable to use said logic in justifying the discovery claim?[26]

However, if we are to claim that there is no monolithic logic of discovery, this would necessarily imply that science does not have a monolithic logic of justification either (Nickles 2000).[27]

Before analyzing the nature of discoveries made by AI, I will note that support for the existence of a logic of discovery fluctuated due to various historical contingencies. One such contingency was a declining belief in the possibility of generating

24. I will revisit the idea of heuristics, and the implications of the same for ML modelling, in section 8.2.
25. Please refer to the previous section where I distinguish between "ad hoc" and "post hoc".
26. Note that the problem of rational reconstruction of discovery claims is similar to the problem of ad hoc (or post hoc) explanations devised to justify the outputs of opaque ML models.
27. This is a key insight and will be used to flesh out the central argument of the thesis in chapters 7 and 8.

infallible knowledge. Although post hoc confirmation could only provide provisional justification, nevertheless, the inability to provide absolute justification was no longer a significant concern. This is because the very idea of absolute justification (like the idea of absolute and certain knowledge) was met with skepticism. As Laudan himself puts it:

> Herschel, Whewell and Comte all acknowledged that there is no formula for producing true theories. As fallibilism emerged, there was an unmistakable shift away from the analysis of genesis towards the post hoc evaluation of theories. It was argued that theories could not be proven to be true and that the most we can expect is that they can be shown to be likely or probable. (Laudan 1981, 188)

5.3.3 Discovery made by AI

Can a machine make a scientific discovery, or are discovery acts exclusively performed by human scientists? It is uncontroversial that machines have aided scientists in performing discoveries. However, there's an ongoing debate as to whether the machines themselves deserve credit with the generation of knowledge or if they merely speed up data processing.

If an AI could sufficiently automate the discovery process, then it is possible that the machine could earn the status of a "co-developer" of knowledge along with humans scientists rather than merely being a tool that is used by scientists (Schickore 2022).

Scientific discovery by AI can be conceptualized as a problem solving exercise within an information processing system. A computational algorithm explores different pathways in a problem space and uses heuristics to efficiently limit the range of possible candidate solutions. However, there are various limitations to these AI programs. It is difficult to identity the relevant data and to properly represent a particular scientific research project as a problem space in a computer. Nevertheless, deep learning methods have renewed interests in the possibility of making discoveries using computer programs (Schickore 2022).

Using ML in making scientific discoveries is also linked to the rising popularity of big data algorithms[28] and how they might modify conventional scientific methods:

> It is also still an open question whether data-intensive science is fundamentally different from traditional research, for instance regarding the status of hypothesis or theory in data-intensive research. (Pietsch 2015, as cited in Schickore 2022)

28. Please refer to Leonelli (2020).

Another salient point from the cognitive perspectives on discovery stresses the importance of understanding human cognition in general, and the discovery process in particular, as a form of *model based reasoning*. Mental procedures lead to the manipulation of mental models of the real world and subsequent formulation of novel models. This raises questions about the possibility of AI-ML programs developing world models of their own.[29]

A central argument in this thesis posits that the adoption of AI will necessitate redefining central concepts in scientific practice. One reason for this shift is due to modifications in the conventional scientific method as practiced by experts using ML techniques. However, another factor behind this shift is due to the fact that AI is emerging as a new, non-human cognizer in scientific practice. For instance, philosophers have suggested that the concept of creativity in discovery should be extended such that machines can also be said to perform creative acts. This has led to philosophers drawing a distinction that can go beyond *anthropological creativity*. Although a machine, being non-human, cannot have anthropological creativity, a machine can still be said to possess *metaphysical creativity*, which is defined as something that can lead to a radically *new* thought or action. Here, the qualification of *new* is defined such that the thought or action is unaccounted for by available knowledge and constitutes a radical break with the past (Schickore 2022).[30]

In this chapter, I described the triad of explanation, prediction, and discovery in ML modelling. In the following chapters, I will focus my attention on the concepts of prediction and discovery as they are used in ML modelling in science. In the next chapter, I will demonstrate a plurality in the usage of the concept of prediction in ML modelling and substaintiate my argument by analyzing a case study. However, before I begin to frame the central argument of the thesis, I will take a small detour to briefly discuss philosophical methods, and explicitly outline those methods that are employed in this thesis to support my arguments.

29. Please refer to Mitchell (2025) for an accessible introduction.

30. Although one can think of sci-fi scenarios involving post singularity AI that might be as creative as any human, it is unclear whether we can consider the current generation of AI models as being creative.

CHAPTER 6

THE CONCEPT OF PREDICTION IN ML MODELLING

6.1 A short detour on methods

In this section, I will discuss the fundamental concept of method in academic philosophy. I will primarily draw from Sundar Sarukkai's work titled 'Philosophy and Method' (Sarukkai 2023).[1] At its most basic, *method* is a set of tasks or rules that is followed to achieve a particular *goal*. An analysis of method is equally illuminated by how it is *practiced*, as well as what it seeks to achieve (the *goal* towards which the method is employed). This general notion of method can also be utilized to understand methods in philosophy. In this thesis, I have employed philosophical methods (the *practice*) to understand scientific-philosophical problems (the *goal*).

Philosophical methods include analysis, creating and clarifying concepts, making and assessing arguments, the use of various techniques from the field of logic, and so on. These methods are used to realize various goals, one of which is to clarify how seemingly similar concepts might be distinct. This is the one of the central goals of this thesis, where I try to show a plurality in the meaning of the concepts of *prediction* and *discovery* in the context of ML modelling in science, and I will describe the various methods that can aid the realization of this goal.

Although doubt is often what leads to the start of an inquiry, unchecked doubt can lead to pessimistic skepticism that can paralyze the progress of any academic project.[2] Therefore, rather than resorting to radical suspicion that can lead to chaotic doubting,

1. Needless to say, I will also be drawing from various discussions, courses, and publications from scholars, my supervisor, expert, and various other academic mentors. An explicit acknowledgement of all sources and persons seems impossible.

2. Sarukkai makes a similar argument in Sarukkai (2015), where he notes that one cannot question absolutely everything while engaging in a discourse or inquiry. An inquiry will have to find some common ground and accept some postulates to be able to make any progress.

philosophers need to employ *methodological* suspicion while dealing with themes of appearance versus reality.

In this thesis, I have also avoided scholasticism and other such methods that might place undue emphasis on older published works, which is where I diverge from other traditions of philosophy (like continental philosophy) that place great value on the methods of hermeneutics and on the commentaries of older established works. However, I use the method of etymology to trace and understand the development of key concepts like "AI", "epistemic opacity", and so on to gain philosophical insights about the contemporary usage of these terms. The methodology in this thesis values knowledge that is imbued with the virtues of simplicity, verifiability, and avoiding grand metaphysical themes that aim to unify a great number of phenomena. These sets of values find common ground with scientific discourses, as put best by Sarukkai:

> Analytical philosophy is not just about the type of problems or the concepts it prefers to use in the act of doing philosophy; it is also based on a belief that language has to have clarity, avoid unnecessary obfuscation, and state things as they are. One can see that the use of language in analytical philosophy is quite similar to that in scientific discourse. (Sarukkai 2023, 100)

At this point, it will be useful to differentiate between *doing* philosophy versus *using* philosophy, and how this distinction maps to the difference between the academic fields of general epistemology versus philosophy of science. Unlike epistemology, the very existence of the field of philosophy of science is premised on the applicability of philosophical methods in understanding science.

To further appreciate this point, I will highlight another distinction between philosophy *of* science versus philosophy *in* science. For Pradeu et al. (2024), philosophy *in* science is a recent trend in the larger domain of philosophy *of* science; where philosophy *of* science is the practice of scholars applying philosophical tools (or using philosophical methods) on the concepts, methods, and other constituents of science in order to address philosophical issues; whereas philosophy *in* science pertains to the practice of philosophers attempting to address scientific questions and make scientific contributions in a particular domain by using philosophical methods.[3] Pradeu et al. (2024) also states that most philosophy *of* science can be understood as philosophy *on* science, in contrast to philosophy *in* science.[4] Even though both of these fields have

3. Pradeu et al. (2024) employ bibliometric tools to collect empirical data on the nature of "philosophy *in* science" publications. They subsequently use this data to come up with three necessary elements for classifying a work under philosophy *in* science. A work in philosophy *in* science should be: 1) addressing a scientific problem, 2) by using philosophical tools, 3) to make a scientific proposal. The authors use this framework to demonstrate how various paradigmatic examples of philosophy *in* science publications can be said to incorporate all the three elements.

4. Yet another way by which philosophy *of* science and philosophy *in* science are distinguished is by noting that philosophy *of* science does not limit itself to answering scientific questions. The discipline of philosophy *of* science also aims to address philosophical questions by drawing from scientific knowledge.

certain common practices, nevertheless, they differ on the basis of the epistemic goals that motivate these practices.

A commonsensical understanding of science would place grave doubts on the capabilities of philosophers in generating scientific knowledge, and the skepticism is warranted. Would philosophers have to publish in science journals in order to make scientific contributions? At the very least, will the scholarly works by philosophers have to be cited by scientists?

In fact, the answer to both of these questions is yes. However, the skeptic can claim that a mere publication in a science journal cannot qualify something as a scientific contribution. In response, Pradeu et al. (2024) defines the *contribution* of a philosophical paper in science as constituting desiderata that go beyond mere *intervention* (in the form of philosophers publishing in science journals) or *visibility* (in the form of philosophers being cited by science journals).

Pradeu et al. (2024) distinguishes between three different categories of philosophers contributing to science: 1) by producing novel scientific results, 2) by producing novel scientific tools, or 3) by participating in a scientific debate. I will focus on 2), and note that philosophers can produce novel scientific tools either by providing (or modifying) methodologies or conceptual tools.[5] One such method is that of providing conceptual clarification of concepts used in science, for instance, by investigating and/or proposing a scientific definition or distinction (Pradeu et al. 2024, 392). In a similar vein to category 2), this thesis aims to contribute to science by targeting an existing methodology or framework. I posit that the extant philosophical framework constitutes a plurality of usage of fundamental scientific concepts and should be modified. I substantiate this claim by analyzing a case study to demonstrating the epistemic utility of drawing a distinction between different conceptual notions that are associated with fundamental concepts in ML modelling.

Philosophers of science often use a case study to substantiate their philosophical arguments. But before we analyze this particular method of supporting an argument, I should highlight the difference between citing a scientific work as a *case study* versus citing a scientific work as a mere example in a philosophical project. When a scientific work is cited as a mere example, it only serves as a brief illustration of a broader philosophical point. The paper is not analyzed in detail, rather, the focus is on the results of the paper. Put simply, when a philosopher is citing a scientific work in a philosophical paper, their focus largely remains on their central philosophical argument rather than the specific details of the cited paper. On the other hand, a cited scientific work becomes a *case study* when it is at the center of philosophical analysis. A philosopher analyzes a case study in greater detail by examining the methods of the scientists, deducing consequences from the scientific results, and drawing out assumptions from

5. Extant or otherwise.

the scientific work.

Currie (2015) provides a detailed defense of the use of a case study methodology in the philosophy of science in a work titled 'Philosophy of science and the curse of the case study'. They first highlight the possible objections against the utility of scientific case studies in philosophical projects by stating that:

> . . . case studies, by their nature, are peculiar and individual, but the generalizations philosophers seek are broad and unitary. (Currie 2015, 554)

Furthermore, they note that:

> And so, because any particular aspect of science is likely to be heterogeneous, and a case study is a single data point, if case studies are inductive bases, we are making a mistake: this is the curse. (Currie 2015, 558)

Currie is framing a critical description of the case study methodology by conceptualizing it as a general problem of finding a representative sample from a heterogeneous population. A philosopher cannot be expected to study all the scientific papers in a given domain. This limits them in having to select a few papers in such a way that their selected sample can be said to appropriately represent the scholarship in a domain. However, if science were to be sufficiently heterogeneous, then the results from a case study might not be useful in providing a more general understanding of a particular scientific domain. While formulating philosophical arguments, there are limitations to the kind of inductive support that can be drawn from using a handful of case studies. Faced with this issue, how are we to proceed?

Currie responds by noting that the utility of case studies can extend beyond providing mere inductive support for a philosophical argument, and posits that case studies can help demonstrate the utility of conceptual tools in explaining and critiquing scientific problems and practice. Even if science were to be heterogeneous, nevertheless, it can still be "patchily unified". There are bound to be some common elements or themes in the methods and concepts which can be used to draw meaningful generalizations over an appropriate scale.

One major advantage of using case studies is that they provide philosophers with a shared empirical basis for philosophical debates. This is because claims from both sides of the debate can be verified by testing them with respect to a common example case from science. Currie emphasizes the point that the discipline of philosophy of science (like any other academic discipline) is pursued as a *social* activity, where debates facilitate progress in solving difficult questions. Philosophical discourse can benefit from using case studies as they provide shared examples that can be used as paradigm cases for verifying philosophical arguments. This allows philosophers to be more

explicit and rigorous while constructing arguments and engaging in debates. This is my justification for a re-analysis of DeVries et al. (2018)'s work that has been cited in Duede (2023); in order to extend Duede's philosophical arguments, while sharing a common scientific work as an example. I will do so in section 6.3, but before that, I will provide a brief review of the concept of prediction as it is used in the context of ML modelling in science.

6.2 Explicating prediction in the context of ML modelling in science.

As we have already seen, the literature on philosophy of ML (Machine Learning) in science lists the epistemic desiderata of modelling like prediction, explanation, understanding, and so on, and describes the connections between them. There are debates about the nature of the trade-offs between various epistemic desiderata, such as the choice between valuing predictive power over explanatory value. Opaque ML models, in particular, seem to be at the risk of prioritizing predictive capabilities over explanatory understanding of the target phenomena (Beisbart 2021). Some scholars go so far as to posit that ML models make "predictions without explanation" (Srećković 2021).

In light of such monumental claims about the implications of adopting ML models in science, it will be helpful to draw out the concept of prediction as used by philosophers and scientists in the context of ML modelling. I will now review a set of scholarly works in the context of ML modelling to motivate the need for an explication of prediction.

There is no shortage of claims that herald major disruptions in scientific methodology with the advent of Artificial Intelligence (AI) technologies (Andrews 2024a). The rising adoption of these AI-ML models has prompted philosophers to explicate the fundamental concepts of epistemic opacity (Facchini and Termine 2022; Beisbart 2021), explanation (Rudin 2019; which is just a singular work from the entire field of XAI[6]), interpretability (Beisbart and Räz 2022), understanding (Fleisher 2022; Sullivan 2022), and so on in the context of ML modelling. However, a central concept lacking explicit treatment in the literature is that of prediction. I will now review a few of these works to demonstrate a lack of clarity (in individual works) and coherence (while comparing different works) in the scholars' usage of prediction.

Srećković et al. (2021) describe a functional relationship between predictions and explanations in science, wherein predictions and explanations feed into each other.

6. eXplainable Artificial Intelligence (XAI) models refer to models that are used to understand other opaque ML models. These models attempt to devise post hoc explanations for the outputs of opaque ML models.

However, the primary role of explanation is to increase the capabilities of scientists in making novel predictions. They further argue that ML has the potential to sever this functional relationship because ML models can make "predictions that are independent of an explanation" (Srećković et al. 2021, 171). They further state – "Obtaining kinds of "free" predictions (not derived from explanatory efforts) would, essentially, disrupt the functional relationship of prediction and explanation commonly encountered in science."

The authors do not provide us with an explicit definition of prediction such that we can assess their claims about the nature of predictions that are independent of explanation. As we shall see later, the problem perpetuates as I attempt to compare Srećković et al.'s usage of predictions with that of other scholars.

Fleisher (2022) attempts to draw similarities between XAI models and scientific models. He says:

> I will argue that XAI methods can similarly be seen as idealized models. The ways they misrepresent the functioning of black box models (like DNNs) are also idealizations (*at least in cases where things go well*). (Fleisher 2022, 548, emphasis added)

Fleisher later expands on what he means by "things going well." While confronted with the question of either accepting or rejecting a newly trained model, Fleisher would direct us to a set of epistemic desiderata that are supposed to guide us in assessing the model. If the model fulfills certain epistemic requirements, then the model should be accepted as a scientific model, regardless of the model's architecture. The argument is extended to include opaque ML models. Opaque ML models might not have perfect similarity to the target phenomena.[7] Nevertheless, an opaque ML model can still be accepted based on "their empirical adequacy, *predictive power*, explanatory power, and theoretical virtue" (Fleisher 2022, 556, emphasis added).

Yet again, we see the central importance that is placed on the predictive capabilities of a model. However, the author does not provide an explicit definition for the same.

In the same vein as Fleisher, Sullivan (2023) argues for treating ML models akin to scientific toy models by drawing parallels between the misrepresentation in ML models and the idealizations of toy models. Not all misrepresentations are idealizations; a misrepresentation is an idealization only if it is a successful misrepresentation. But this begs the question – how do we know that the end product of an ML modelling exercise is a success? How do we know that the ML model has actually captured a "successful idealization"? Sullivan answers by noting that:

7. As Fleisher argues, this is something that is true for a lot of other scientific models that idealize the target phenomena.

> In general, idealizations can be successful empirically if they have *predictive power* (Mizrahi 2012) or are safe for engineering use (Batterman and Rice 2014; see Lawler 2021). On this score, ML models may do well because of their *high predictive power and usefulness*. (Sullivan 2023, 9, emphasis added)

An explication of the concept of prediction would help us understand the meaning of "predictive power," which seems to be the primary factor in determining a model's "usefulness" for Sullivan.

In another work, Sullivan (2022) talks about the various kinds of understanding that ML models can provide. She claims that:

> . . . many DNN models address simple classification tasks, like identifying a number from a handwritten note. One could reasonably argue that there are no explanatory questions one could ask of such a model; only *mere prediction is possible*. Maybe so. What I have been arguing for in this article is that the complexity and black box nature of DNN models does not prevent understanding of phenomena. (Sullivan 2022, 129, emphasis added)

It is not clear to me whether the kind of predictions that Sullivan refers to as "mere predictions" are similar to the kind of predictions that Srećković et al. (2021) refer to as "predictions independent of explanations."

Duede (2023) advocates for the usage of opaque Machine Learning (ML) models in scientific practice by invoking the *context distinction* in scientific discovery.[8] He situates ML models as part of the context of discovery and posits that scientists are justified in using ML models as heuristic exploratory tools despite their opacity. He substantiates his argument by analysing a case study by DeVries et al. (2018) wherein an ML model is used as an aid in the discovery process.

A corollary from Duede's thesis is that ML models can be smoothly integrated into the standard scientific methodology of various disciplines in science.[9] According to Duede, these ML models can aid scientists in the discovery process and help them in making novel predictions. In response to Duede, I will argue that the philosophy of ML modelling in science is interchangeably using different notions of discovery and prediction which is incongruent with the conventional usage of these concepts in conventional scientific methodology.

As I have already presented in chapter 6, the context distinction in discovery is conceptualised in various ways by scholars (Schickore 2022). However, in most cases, discovery is seen as a process rather than a singular event. Therefore, it is not immediately clear as to when an inquiry results in a discovery and where are we to draw

8. The distinction between the context of discovery and the context of justification (Schickore 2022).

9. I will bracket discussions on the problems associated with treating science as comprised of a monolithic method of practice. For now, we can restrict our argument to a particular discipline.

the context distinction in a discovery process. In the next subsection, I will analyze the ML modelling pipeline with the perspective of discovery as a drawn out process that requires justification at the end.

6.3 Analyzing a case study

In a paper titled 'Deep Learning Opacity in Scientific Discovery', Duede (2023) highlights the mismatch between philosophical pessimism and scientific optimism in the philosophy of ML modelling in science. To justify using opaque ML models in making scientific discoveries, Duede frames the discussion as part of a wider discovery process. Duede argues for situating these models in the context of discovery which seems to automatically absolve them of requiring justification. Duede states that ML models[10] can be used by scientists in the context of discovery because the discovery process does not necessarily require interpretability for its justification.

Duede (2023) justifies the use of an opaque ML model in DeVries et al. (2018)'s work by noting how the ML model realised greater predictive[11] capabilities compared to extant theory. This led scientists to *discover* the significance of hitherto overlooked parameters in extant theory. Because the ML model was able to demonstrate greater predictive capabilities, the scientists had good reason to try to apply post hoc (or ad hoc) theorization on the outputs of the ML model, which finally led to the discovery of a novel hypothesis.

However, the ML modelling pipeline that is described by Duede involves at least 2 different notions of both prediction and discovery. These are the discovery and predictions of the ML model versus the discovery and predictions of the revised theory at the end of the modelling pipeline.[12] Because we operate in the larger consequentialist paradigm in the philosophy of discovery, scientists do not accept a discovery claim just because it is the result of a reliable process. In the process of accepting a discovery claim, the emphasis is on assessing the downstream consequences of accepting the discovery claim. The acceptance of these downstream consequences justifies the discovery claim.[13]

Duede is right in pointing out that philosophers have not kept up with scientists who are able to successfully employ these ML algorithms in various domains. Nevertheless, philosophical skepticism is warranted, because these ML models do represent a distinct methodology of conducting science owing to a different *culture of prediction* surrounding it. This new culture of prediction places great emphasis on realising predictive capabilities even if this comes at the cost of developing opaque models that

10. Duede is arguing for a subset of ML modelling exercises. Some ML modelling exercises do aim to cross over to the context of justification proper.
11. I will later show how the usage of the concept of prediction here is ambiguous.
12. I will demonstrate how these are different by detailing the cited case study in the next section.
13. Please refer to section 6.3.2 for a more detailed discussion on the topic.

are not amenable to scientific understanding (Srećković, Berber, and Filipović 2022; Johnson and Lenhard 2024).[14]

Duede acknowledges the epistemic virtues gained by scientists when they gain understanding of a model. However, he also notes that there are cases where the opacity of a model is irrelevant to justifying the claims of a ML model. The opacity of a ML model is only relevant in situations where the outputs of ML models are treated as candidates for scientific knowledge in their own right. The outputs of these models are only used to formulate hypotheses, and these hypotheses are further tested, modified, and pursued independent of the history of how these hypotheses was formulated.

The end product of the discovery process will require rigorous justification. However, according to Duede, the part played by an opaque model in the process can be insulated from the same rigor of justification. This is similar to how the leaps of imagination and insight in a scientist's mind that initiate a discovery process need not (and possibly cannot) be rigorously justified. As Duede himself states:

> The outputs of neural networks can be used to guide attention and scientific intuition toward more promising hypotheses but do not, themselves, stand in need of justification. (Duede 2023, 1094)

Scientists can apply these ML algorithms on a wide range of data sets and develop models that can outperform theory based models in terms of predictive capabilities. However, this comes at a cost. Unlike theory based models, an ML model does not have any theoretical constraints. This allows the ML model to (potentially) violate widely accepted and valuable general principles of the domain. For example, an ML model could violate conservation laws and continuity assumptions in order to gain a marginal increase in exactly predicting the evolution of a physical system. The architecture of these ML models prioritizes predictive capabilities over all other epistemic desiderata like explanatory value, scientific understanding, and so on.[15]

To substantiate his claim, Duede takes a case study by DeVries et al. (2018) wherein scientists are able to train a ML model to successfully predict aftershock patterns from a data set of the main shock patterns of earthquakes. This by itself is impressive; however, the scientists were subsequently able to devise a post hoc explanation to account for the predictive success of the ML model. In this case, the novel theoretical insight was in understanding the significance of hitherto unrelated and seemingly insignificant geophysical parameters. The significance of these parameters is then injected into the extant theory. This is the "reworking" of extant theory that Duede mentions in his work.

14. I will revisit the idea of a culture of prediction in the final section of this chapter.

15. Not all AI-ML models are like this. Some scholars advocate for exclusively constructing models that are interpretable to experts (Rudin 2019).

I will argue that the *post hoc explanation* that scientists provide in an attempt to explain the opaque ML model has to satisfy some additional criterias. This is because scientists use these models for epistemic desiderata other than just making predictions. Scientists also use modelling to gain an understanding of the phenomena they are studying and to further develop theory. The post hoc explanation, or the post hoc hypothesis, needs to be epistemically transparent, integrable into theory, and amenable to scientific understanding and manipulation. These conditions happened to have been satisfied by the post hoc explanation in DeVries et al. (2018)'s case study. However, I will argue that if the scientists could not provide such a post hoc explanation, then it would be very difficult to assess the epistemic status of the opaque ML model's outputs.

I think that DeVries et al. (2018)'s work has gained the recognition that it has only because the scientists could successfully provide a post hoc explanation for the outputs of the ML model. But what if the scientists could not provide such an explanation? In this case, the outputs of the opaque ML model in isolation (that is, without the associated ad hoc explanation) might not have been recognized as a discovery claim.[16]

Duede is correct in noting that the novel hypothesis that is formulated by an opaque ML model is not implicated in, nor relevant to, justifying the reworked theory. However, I will argue that it is only when the reworked theory is able to explain the outputs of the ML model that these outputs are justified, and consequently, awarded their discovery status. When a scientist finds herself in the middle of the discovery process, there is no guarantee that the outputs of the ML model will be able to make modifications to extant theory. This raises questions on the epistemic status of the outputs of these opaque ML models in isolation from any form of theoretical justification other than their predictive capabilities.[17]

In the following section, I will present my own analysis of DeVries et al. (2018)'s case study and contrast my arguments with those of Duede (2023). I will do this by demonstrating a plurality in the usage of the concept of prediction and discovery by Duede.

6.4 Different notions of prediction and discovery

In the case study by DeVries et al. (2018), scientists were studying the distribution of the results of the ML model compared to theory and were surprised to find the probability distribution that was assigned by theory was largely uncorrelated with the results of the ML model. This was a brazen challenge to expert intuition and theoretical predictions. It could either mean that the modelling exercise had unearthed a deeper

16. Even if they were to be accepted as a discovery claim, their epistemic value would be significantly lower. I will flesh out this argument in the next section.

17. The issue becomes more tricky when the theory is incompatible with the results of the ML model.

problem with extant theory, or that the ML model had made a mistake. In this case, the ML model was vindicated when it was able to guide scientists to improve the extant theory.

In this particular scientific problem of predicting aftershock patterns, the ML model was able to demonstrate greater predictive capabilities compared to extant theory. Moreover, the probability distribution of the extant theory was very different from the ML model. This motivated scientists to modify their theory such that the reworked theory would make predictions that were similar to that of the ML model.[18] Throughout the discussion, Duede assumes that the epistemic assessment of the reworked theory can be performed independent of the ML model that has inspired modification to the theory. This makes considerations of opacity of the ML model irrelevant in the context of justification.

But I think that there are two different notions of predictions at play here. This is the difference between a prediction versus an *empirical consequence* of a model. Duede himself seems to implicitly make this distinction. While talking about why the predictions of the ML model need not be justified, he states:

> This is because it is not the network's *predictions* that stand in need of justification but, rather, the theory's itself. (Duede 2023, 1097, emphasis added)

Duede is assuming that the predictions of the network[19] are of a different nature compared to the theory's prediction. Perhaps a better term to use will be that of *predicting* rather than prediction. This is because the emphasis is not so much on the results of a predictive practice, as much as on the system that makes the prediction and the justification that is given for said prediction. Nevertheless, the two different notions of prediction here are:

1) The *outputs* of opaque ML models, which I will call prediction$_1$ as opposed to
2) The conventional notion of prediction as the *empirical consequences* of scientific theories or models, which I will call prediction$_2$.[20]

Unlike prediction$_2$, prediction$_1$ lacks coherence with theory. Moreover, prediction$_1$ also differs on the basis of the elements that constitute the predictive system. Traditionally, scientists have relied on models with interpretable elements[21], whereas the opaque ML models generating prediction$_1$ lack this property. After making this implicit

18. Which would also increase the predictive capabilities of the theory.
19. The network here refers to the neural network architecture behind the ML model.
20. This is similar to Forster (2014)'s definition of scientific predictions. Please refer to section 6.2 for a more detailed discussion.
21. Please refer to Freiesleben et al. (2024) for a discussion on the implications of using opaque ML model constituing uninterpretable elements, and its implications on the practice of science.

distinction, Duede describes the reworked theory in DeVries et al. (2018)'s case study and goes on to add:

> The reworked theory is justified in ways that are consistent with the norms of the discipline—it relates known geophysical properties in ways that are consistent with first principles, it aids in the explanation and understanding of aftershock dynamics, and it outperforms extant theory in prediction. (Duede 2023, 1097)

I will argue that Duede is able to use the term predictions to describe the outputs of the ML model in hindsight only because the reworked theory was able to achieve all the epistemic desiderata listed by Duede.[22] I think that the outputs (prediction_1) are accepted as scientific discovery if and only if said outputs are subsequently understood by scientists in a manner such that these outputs can be integrated into extant theory (prediction_2). If such an understanding proves to be illusive, then the epistemic status of the outputs of an opaque ML model (prediction_1) is, at best, uncertain. In the case study taken by Duede, the numerical results of both prediction_1 and prediction_2 happened to closely converge towards the end of the modelling exercise. However, these two notions of predictions are not the same, and they need not necessarily converge in all modelling exercises.

These results will challenge Duede's central thesis where he states that opaque ML models can be effectively used for genuine discovery and deeper theoretical understanding in science (Duede 2023, 1097). Duede defends this thesis by noting that the formulation of the hypothesis[23] is entirely situated in the context of discovery. However, as I have attempted to demonstrate in this section, the subsequent pursuit[24] which justifies the novel discovery claim of the ML model cannot be divorced from theoretical justification.

To better understand the last point, I will contrast the epistemic status of the outputs of a ML model with the results of an experiment. Because experiments enjoy a higher epistemic status in scientific practice, an inconsistency between the results of an experiment and the predictions of a theory can justifiably lead to modifications to the theory. The predictive capabilities of these ML models are raising questions as to whether we can afford to provide the same epistemic status to the outputs of opaque ML models.[25]

22. As an aside, i am not sure if the reworked theory has aided in explaining the aftershock dynamics or has merely been used to make better predictions. This is because the reworked theory includes ad hoc hypothesis produced by the ML model which are very unintuitive.

23. Which, in this case, is the output of the opaque ML model.

24. Please refer to the idea of the context of pursuit in section 6.3. More importantly, I will analyze the different modes of justification in the context of pursuit in the next chapter as I divert my attention from predictions to discovery.

25. This question also arose with the advent of computer simulations in science. However, unlike

At least in DeVries et al. (2018)'s case, the scientists were open to modifying the theory in response to the outputs of the ML model. The willingness to include opaque ML models in the discovery process seems to conform the emergence of a new *culture of prediction* surrounding our current scientific paradigm (Johnson and Lenhard 2024).[26] In this new culture of prediction, scientists are open to the possibility of making major modifications to their theory on the basis of opaque models that they do not understand. Similarly, philosophers like Duede say that *inductive considerations*[27] are sufficient to establish reliability in cases where the ML models are leading scientists to major breakthroughs (Duede 2023, 1097).

I will now attempt to tease out two different notions of predictions and discovery that are being interchangably used in the discussion. The opaque ML model tracked aftershock patterns with an accuracy of AUC[28] 0.849 (DeVries et al. 2018). I will call the outputs of this ML model as prediction$_1$. On the other hand, the predictions that result from the reworked theory seem to be of a different nature altogether, and I'll call these prediction$_2$. While reading the original paper cited by Duede, it is clear that this distinction is useful, as the scientists note how prediction$_2$ accounts for 98 percent of the variance of the outputs of the ML model. Or to put it differently:

$$\text{AUC prediction}_2 < \text{AUC prediction}_1 = 0.849$$

Even though prediction$_2$ is less accurate than prediction$_1$, prediction$_2$ has a higher epistemic status for scientists due to its explanatory value and ability to integrate with extant theory.

Is it appropriate to use the same word "prediction" to refer to the *outputs* of DeVries et al. (2018)'s opaque ML model as well as the *theoretical predictions* of the reworked theory? A standard account of predictions in philosophy of science defines predictions as the empirical consequences of a scientific model (Forster 2014). This leads us to question whether we can qualify an opaque ML model as a scientific model? Should we reserve the concept of prediction to only include the empirical consequences of scientific theory/models proper? Or can we also use the concept of prediction to include the *candidates for scientific knowledge*, like the outputs of an opaque ML model? (Duede 2023, 1093)

Like predictions, there is a plurality in the usage of the concept of discovery as well. In the process of creating a ML model that could successfully outperform the predictions of the extant theory, DeVries et al. (2018) made a discovery. I will call this

the data-driven approaches of ML architectures, computer simulations use "top-down" methods of inferences that flow down from theory and are qualitatively different from the algorithms behind AI-ML models (Winsberg 2022).

26. I will revisit the idea of a new culture of prediction in more detail in the next section.

27. More work on defining the concept of predictions as opposed to the general idea of *inductive considerations* will be illuminating here.

28. Area Under the Curve.

as discovery$_{E1}$. Here, $E1$ is the epistemic value of the discovery claim.[29] This discovery led to the set of outputs that I am referring to as prediction$_1$. Subsequently, the authors were also able to closely reproduce the predictive capabilities of the opaque ML by modifying the extant theory. The predictions of this reworked theory are termed as prediction$_2$. This latter discovery claim which comprises scientists' understanding of the significance of a few geophysical parameters will be labelled as discovery$_{E2}$. The noteworthy point here is that discovery$_{E2}$ did not merely provide scientists with an increase in predictive capabilities (like discovery$_{E1}$), but also with greater theoretical understanding (unlike discovery$_{E1}$).

With the terminology set, I will argue that the utility behind distinguishing these two notions of discovery is due to the significantly greater epistemic value of discovery$_{E2}$ over discovery$_{E1}$. The discovery claim of the ML model gained more epistemic support after it was found that extant theory could be reworked to reproduce the outputs of the ML model. This raised the epistemic value of the initial discovery claim. This gain of epistemic support seems to pass unacknowledged when the same word is being used to identify the outputs of ML models (prediction$_1$) and the results of the reworked theory (prediction$_2$). One simple way to highlight the significance of this distinction is to imagine the epistemic value of the output of an ML model if it could not be integrated into extant theory. If this were the case, the output of the ML model would only exist as an isolated result with very little epistemic value; and DeVries et al. (2018)'s work would probably not gain the same amount of recognition that it has now.

6.5 A new culture of prediction

Recent work in the history and philosophy of prediction further complicates[30] the fundamental concept of prediction in ML modelling. As I have already noted in subsection 6.2.5, Johnson and Lenhard (2024) provide a descriptive account of 4 broad cultures of prediction in the history of science. These are the *rational* and the *empirical* cultures. An *iterative-numerical* culture which emerged when computers began to be used to understand a handful of models, and fourthly, an *exploratory-iterative* culture wherein the creation of large data sets and easy access to computational resources allowed scientists to explore a wide range of model spaces, for example, using the method of parameterization.

In the concluding chapter, Johnson and Lenhard (2024) hint at an emerging fifth culture of prediction surrounding the use of AI-ML modelling in science. This new culture uses opaque ML models to try to obtain *pure predictions*[31], even if it comes at the

29. The epistemic value of a claim is a function of the epistemic desiderata that it provides.
30. Illuminates?
31. One account of prediction describes the concept as a logical link between hypothesis/theories

cost of other epistemic desiderata like explanatory value and scientific understanding. With time, these ML models might gain footholds in ever more disciplines of science, where they will make predictions with no regard for general principles of extant theory. Due to their opacity, it will be very difficult to assess which set of background hypothesis is responsible for the model's predictions.[32] Examining Duede's work, and the work of other scholars in the philosophy of ML modelling in science, will help us in responding to novelty claims about a the culture of prediction surrounding ML modelling. Any inquiries on this front will benefit from:

1) Clarifying the notions of *prediction* and *discovery* in the philosophy of ML modelling in science. This will involve drawing contrasts between the topical notions of prediction and discovery, and how these concepts have been understood in the conventional accounts in philosophy of science.
2) Assessing the epistemic status of the outputs of opaque ML model that exist in isolation from any form of post hoc theoretical justification.
3) Determining the relationship between, and priority of, the competing epistemic desiderata of prediction, explanatory value, scientific understanding and so on in the context of philosophy of ML modelling in science.

If these AI models continue to improve in their capabilities, then one could argue for the possibility of these models to go beyond making mere predictions. These models could gain the ability in formulating incommensurable theory that is superior to human theory. While discussing the concept of the *underdetermination of theory by data*, Newton-Smith (2000) takes up the example of an alien civilisation whose evolution of scientific theories took a different historical and epistemological trajectory than ours. This idea can be extended to ML and other AI models having a different nature and taking a different trajectory in the evolution of their scientific theories than humanly constructed scientific theories. Therefore, assessing the epistemic status of these opaque models is an important task for philosophy of science.

In the next chapter, I will switch my attention from the concept of prediction to the concept of discovery in ML modelling. I will focus on approaches in understanding the machine discovery process that emphasizes the development and preliminary evaluation of promising hypotheses prior to rigorous testing. I will bracket the discussion

and their observations (Forster 2014). Drawing from these different accounts of predictions will help us understand the idea of *pure predictions* in ML modelling. These pure predictions, divorced as they are from theory, cannot be understood in the same way as conventional predictions in science. The internal elements of a theoretical model are amenable to epistemic access which allows scientists to understand which hypothesis/element of the model is targetted in a particular act of confirmation with an experiment. However, it can be very challenging to understand which part of an opaque ML model is being confirmed in a successful confirmation test of the model.

32. Holism in theory testing is a well recognised problem in philosophy of science which will get unimaginably complicated if opaque ML models dominate the landscape.

on the actual generation of the novel hypothesis or idea in the mind of the research scientist (or the machine involved in the discovery process).

CHAPTER 7

THE CONCEPTS OF DISCOVERY AND JUSTIFICATION IN ML MODELLING

7.1 Introduction

Before I analyze the concept of discovery in ML modelling, I will explicate the more fundamental concept of *justification*. I will motivate the need in paying such close attention to this concept by highlighting the process by which a discovery claim is accepted as a scientific discovery. A discovery claim is accepted as such if a scientist can provide sufficient *justification* for it. This leads us to question: what is justification in scientific discovery, and what determines the adequacy of a particular form of justification?

To *justify* something is to provide rational support for it. We justify a claim (or a hypothesis/idea) by providing rational support that counts as evidence for an argument regarding the truth value of the claim. In science, we can justify a claim by testing the empirical consequences of accepting the claim, or by assessing coherence of the individual claim with existing theory in a domain.[1]

Philosophers of discovery differentiate between different modes of justification in the discovery process (Schickore 2022). In the context of discovery, the process of formulating a claim is subjected to various "weak" forms of justification. These weaker forms of justification are in contrast to "strong" forms of justification, like *consequentialist* and *generativist* justification, that are exemplary of scientific practice. Justification of the weak form can be provided by preliminary appraisal of the discovery claim, demonstrating the discoverability of the claim, and other factors like explanatory power, aesthetic value, or even the pragmatic value of the claim.

1. There can be other avenues of justification, for instance, a theory can be justified on the basis of aesthetic judgements on its simplicity, consensus across disciplines, or even from pragmatic judgements of the theory's potential for generating new research questions.

Given these different notions of justification in science, which ones are used (or can be used) by scientists in justifying the outputs of opaque ML models? What is the epistemic status of the outputs of an opaque ML model before they can be justified using strong modes of justification and integrated into extant theory? One way to conceptualize these outputs is to look at them as provisionally accepted claims or *heuristics*. However, one major difference between conventional heuristics used in scientific discovery and the outputs of opaque ML models is that the latter are in possession of a singular mode of justification.

Due to their opacity, the outputs of an ML model cannot directly be integrated into existing theory. This means that these outputs lack various other modes of justification like their coherence with extant theory. This issue poses significant challenges for scientists working with ML models. In response, I will argue that the claims of ML models are of a different nature compared to the claims of conventional science. This is due to the ML model's lack of representational and theoretical substance, which makes it difficult for scientists to integrate ML claims into extant theory. This undermines the scientists' epistemic trust in opaque ML models (and their outputs) even if these models have other epistemic virtues like their predictive power.

7.1.1 Predictions as justified claims

Difficulties compound when we try to compare the predictions of ML models with those of theory. This is because the nature of predictions of ML models is different from those of conventional, theoretical predictions. The data-driven predictions result from opaque ML models comprising epistemic elements that resist semantic interpretation by experts. Although both the ML model and extant theory make claims about the empirical world, the *process of formulating* these claims are very different.

To demonstrate that these data-driven ML predictions are of a different nature compared to conventional predictions, I should first note how a prediction is different from a mere claim. A *claim* is a statement that asserts the truth value of a proposition and does not necessarily require any form of justification. However, a prediction has to have some form of justification. In this vein, an acceptable prediction warrants justification for its truth value. And the justification for a prediction claim is precisely the fact that the prediction claim is the result of an epistemically reliable *predictive process*.[2] A predictive process can be a modelling exercise or deductions from extant theory, and the outputs of a predictive process can be inferences drawn from theory,

2. Because I am demarcating *claims* and *predictions*, I will argue that even the act of testing theoretically formulated predictions with emperical data is a part of the process of formulating a prediction. The test is a part of the conventional HD model and informs the construction of future predictions. *In this specific context*, the very fact that a test influences the formulation of a prediction allows me to situate post hoc prediction testing as part of the predictive process. This is not true for *claims*, which need not be influenced by any sort of test results.

the outputs of a ML model, the judgement of a domain expert, and so on. Note that all of these predictive processes rely on different *modes of justification*. I will use this insight to argue that one important criterion which allows us to distinguish between different types of predictions is the Hence, if we find that a set of predictions have a different mode of justification, we can also claim that those predictions are of a different nature altogether. The motivation behind developing this vocabulary is to respond to novel scientific-philosophical problems arising in the context of the philosophy of ML modelling in science.

A short note on my definitions: a *form of justification* refers to a fundamentally distinct way in which a prediction is justified and seperates itself from a mere claim. But because the thesis is primarily looking at ML modelling in science, I will use a more specific concept of a *mode of justification* which describes the specific way that a justification manifests in a given *form*. For instance, in a *form* of justification for making predictions using ML modelling, different *modes* of justification can include predictive capabilities, theoretical coherence, generalizability, and so on. While *forms* distinguish between broad types of justification, *modes* capture variations within a particular *form*. In the next section, I will analyze the different *modes* of justification that are available in ML modelling.

7.2 Different modes of justification in ML modelling

Justification for a claim does not solely rely on emperical tests of that claim. If the claim had the additional epistemic virtue of cohering with existing theory, that too would count as valid justification for the claim's truth. However, in the practice of using ML modelling in making predictions, scientists often do not possess this additional mode of justification. Experts are required to justify the outputs of their opaque ML models by relying on their predictive capabilities, even if these models blatantly violate established principles of existing theory.[3] This leads to a trade-off between two different modes of justification that I already mentioned: *empirical adequacy*, and *theoretical coherence*.

As I have already mentioned in section 6.3, *consequentialist* justification is provided by comparing the consequences of accepting a claim with observations. To assess the justification behind asserting a prediction, a consequentialist will test the novel predictions that are the empirical consequence of accepting the prediction claim (Schickore 2022). In ML modelling, the novel predictions can be the predictions that the ML model makes beyond that of its training data set. As I have summarized in section 7.2, Duede (2023) posits that ML algorithms can be justifiably used to formulate hypothesis in the context of discovery. And because the model's outputs are justified on the basis of

3. Please refer to section 7.4 for various examples.

consequentialist justification, the process of how the discovery claim was formulated is irrelevant.

However, a discovery is labelled as such only after the discovery claim has been justified. In the study by DeVries et al. (2018), the outputs of the ML model were accepted as a discovery claim only after they were integrated with extant theory. Moreover, it was the theory itself that was modified (albeit slightly) to cohere with these claims. The scientists found some correlations in the data set using an opaque ML model, and although these correlations were able to make accurate predictions, they did not exactly cohere with extant theory. I summarize this process in order to pose the following question: were DeVries et al. (2018) correct in modifing the theory, or should they have further scrutinized the ML model? I will present a framework to answer this question by extending Duede's argument where I will specify the mode of justification in similar cases of ML modelling in science.

Two different notions of justification exist in the discovery process. The justification that happens *after* the hypothesis has been formulated, and justification for the very *process which led* to the formulation of a hypothesis. Following Schickore (2022), I will call these two modes[4] of justification as *consequentialist* and *generative* justification. These are the two "strong" forms of justification. As opposed to the strong forms of justification, the "weak" forms of justification are a set of methods which assess the value of a hypothesis (like the output of an ML model) during the discovery process itself, and prior to rigorous testing involving "strong" forms of justification (Schickore 2022).

Duede (2023) notes how scientists use a method where they *rationally reconstruct* a discovery claim. This is a mode of "weak" justification that underlies DeVries et al. (2018)'s work where it is suggested that if the scientists had known the significance of using the 4 parameters of existing theory beforehand, they would have made the discovery independent of the ML model.[5] However, is this an acceptable mode of justification, or is it suspiciously similar to a form of ad hoc theorization?

Along with rational reconstruction, there is another "weak" mode of justification called *discoverability* that can benefit our understanding of the practice of post hoc justification for the outputs of ML models (Schickore 2022). As a mode of justification, *discoverability* states that demonstrating the discoverability of new results from existing theory can serve as a justification for the new results. Post-hoc, the scientists can demonstrate how the discovery claim could have been derived from existing knowledge (including empirical data, theory, and even expertise) of a domain. DeVries et al. (2018)'s results can also be said to rely on a mode of discoverability justification.

4. My usage of a *mode* of justification as defined in the previous section is narrower than that of Schickore (2022). Here, I am quoting Schickore (2022)'s usage.

5. Please refer to section 7.4 for a more detailed analysis of the case study.

7.2.1 Heuristics

Other than *discoverability* and *rational reconstruction*, another popular mode of weak justification is the use of *heuristics* in science (Schickore 2022). In the context of ML modelling, this naturally leads us to questions regarding the nature of heuristics in ML modelling, and how they relate to our understanding of heuristics in conventional scientific practice.

In his article titled 'Many Meanings of "Heuristic" ', Chow (2015) seems dismissive of giving any special significance to *computational* heuristics. This is because the heuristics literature largely seems to focus on heuristics as they are understood in cognitive science and philosophy of mind (Chow 2015; Schickore 2022). This makes it difficult to assess the epistemic status of computational heuristics by drawing from the published scholarship.

But before I begin my review of *computational heuristics*, we should first clarify our understanding of *heuristics* as a general concept. Chow (2015) provides us with a negative definition of *heuristics* by contrasting them with Guaranteed Correct Outcome (GOC) procedures. These GOC procedures can efficiently come to a tentative, "good-enough" answer. But more often than not, these GOC procedures do not serve as reliable ways of arriving at the actual answer for a specified problem. Unlike GOC procedures, heuristics do not guarantee correct outcomes. The value of a heuristic lies in the additional epistemic benefits they can provide to a cognizer who is engaged in a problem-solving exercise.[6]

The difference between the nature of heuristics used by humans versus the *computational heuristics* of an AI lies in the representational content of the heuristics. Any human-usable heuristic has to have some perceptual or cognitive representational content. But a heuristic for an AI need not satisfy this requirement. As Chow himself puts it:

> . . . computational heuristics are still heuristics, they are just of a sort different from cognitive heuristics. (Chow 2015, 996)

Defining heuristics in this manner might exclude the internal elements (like parameters) and outputs of AI-ML models that work on a connectionist learning framework

6. Are ML models more similar to GOC procesdures or to computational heuristics? Like heuristics, these ML models also provide additional epistemic virtues to the expert who is employing these models on their data. On the other hand, a ML model, like a neural network, can also mathematically guarantee a solution, however, it might not be the actual solution. Gradient descent is one such commonly used algorithm that can minimize error in a problem statement. A gradient descent method, like other regression methods, minimizes the error of a function by iteratively changing the value of the parameters of the loss function in a ML model. This minimizes the loss function, and produces a function on the training data set that is able to make accurate predictions.

and do not possess any representational content owing to their epistemic opacity.[7]

Chow (2015, 998) further distinguishes between:

1) *Inferential* heuristics, which he also labels as being epistemically opaque, and,
2) *Methodological* heuristics, like some of the more standard and transparent mathematical formulations that are used in understanding scientific phenomena.

This distinction will help us in avoiding the trivialization of the concept of opacity. It is true that understanding comes in degrees, and that any scientist will not have a perfect understanding of a mathematical framework that they are applying in their practice. However, I wish to reserve the tag of epistemic opacity for those models where we do not have any access to the representational content of the epistemic elements of a model. Although both *inferential* and *methodological* heuristics are opaque in the sense that we do not understand *why* they work, but in the case of a standard mathematical framework, we at least have an understanding of *how* it works and *what* it is.[8] These epistemic questions of understanding the *how* and the *what* are separate from the more fundamental philosophical question of *why* these heuristics are successful.

For Chow (2015), *heuristics* are cognitive procedures that can be expressed as rules that one reasons in accordance with.[9] Heuristic procedures do not aim to meet the epistemic standards of theory; instead, they aim to meet standards that are in some sense "good enough" relative to the agent (Chow 2015, 1003). For our purposes, the agent is a scientist working with an opaque ML model. And the immediate concern is about identifying the factors that influence scientists' judgement about the outputs of ML models[10] being *good enough*, and whether scientists are content in solely using these heuristics for their predictive capabilities.

We can build on Chow (2015)'s work and extend his definition of a computational heuristic by starting with a more fundamental notion of heuristic. A *heuristic* is a set of rules or procedures that can suitably be applied in local contexts to get a possible answer to a problem. However, there is no guarantee that a possible answer given by a heuristic is the correct answer. Therefore, whenever a scientist is working with a heuristic, there is this notion of a trade-off between accepting local pragmatic value versus attempting to obtain more global principles. In section 7.4 of this chapter, I will demonstrate how these epistemic trade-offs emerge in different forms across various case studies employing ML models—motivating the need for clarifying the meaning of the concepts comprising these epistemic trade-offs.

7. The analysis can benefit from distinguishing between representational and non-representational reasoning. However, this is beyond the scope of the thesis.

8. Please refer to the example of the simulation of a simple pendulum in Alvarado (2021) as it has been discussed in section 4.6.

9. Chow (2015) brackets the discussion on computational heuristics by restricting his definition to cognitive processes.

10. Assuming that we can treat the outputs of ML models as heuristics.

I will summarize this discussion on heuristics by emphasizing that the process of formulating a hypothesis also involves various forms of heuristic appraisals regarding its epistemic value (Schickore 2022). In ML modelling, the outputs of ML models are candidates for subsequent pursuit as potential discovery claims. Some of the criteria for appraisal are the predictive capabilities of the hypothesis and the coherence of the hypothesis with extant theory. I think that there is an urgnet need for scholars to describe how scientists deliberate on this appraisal process in the context of discovery, especially in the case of ML models that are able to make accurate predictions but are inconsistent with extant theory of a discipline.

Before I present a review of various case studies employing ML modelling, I will address a skeptical argument regarding the philosophical novelty of ML modelling in science.

7.3 Philosophical novelty of epistemic opacity in ML modelling

There are similarities in the process of preliminary appraisal of the claims made by ML models and computer simulation models. Tal (2011) describes one such relatively recent methodology of using computer simulation signatures to establish evidential standards in physics. A computer simulation model that is built on extant theory is used to simulate the outputs of an idealized experimental setup. And if these *simulated signatures* can be explained by the existence of the phenomena under study, this raises epistemic trust in the existence of the phenomena. If these simulated signatures are confirmed by other independent criteria[11], scientists gain further epistemic trust in the existence of the phenomena. In this manner, these "retro-claims" about the existence of a phenomena using simulated data (or outputs of opaque AI algorithms) could be an interestingly novel methodological category of discovery for philosophers of science.

There are some similarities between ML modelling and the use of computer simulations. Both of them are instances where an epistemically opaque technology that lacks sufficient semantic interpretation is being used by scientists. It is almost as if scientists are using technological tools that they do not completely understand, yet, these tools end up being very effective in helping scientists conduct science. This leads us to question whether the issue of epistemic opacity surrounding ML models is simply a new installment of an old problem.[12] Wigner described the "unreasonable effectiveness of mathematics in science" (Wigner 1960) wherein he marvels at the apparent mystery

11. In the case study taken up by Tal (2011), the availability of the results of two different conventional experiments led scientists to infer the existence of the phenomena.

12. Thanks to Varun Bhatta for raising this objection.

behind the successful application of mathematical frameworks in science. Is this similar to the "mystery" of ML programs successfully aiding scientific discoveries?

I can generalize the previous cases in the following manner. In the course of her research, a scientist uses various things that she does not completely understand. Be it the ML functions constructed as a result of training a ML model, or the Hermitian operators in quantum mechanics. All of these are specific categories of mathematical entities that are mysteriously successful in science, and we do not know *why* that is the case.

However, drawing from Sarukkai (2005)'s reply to Wigner, the apparent "mystery" can be extended to the unreasonable effectiveness of successfully applying language in describing the natural world as well. This generalizes Wigner's mystery to the central epistemological theme of "representations versus reality" that can be found in all philosophical traditions.[13] As helpful as it is to situate the problem of epistemic opacity of ML modelling in these long standing and richly debated traditional themes of philosophy, I think that equating the opacity of ML modelling to that of the mystery of using mathematics or language to describe the world risks trivializing the issue.

My emphasis on why opacity is a novel scientific-philosophical problem is that we have little to no semantic interpretation of the epistemic elements of these ML models. Despite the lingering mystery surrounding the applicability of Hermitian operators and Hilbert spaces in the context of the theory of quantum mechanics, scientists nevertheless have an understanding of *what* the mathematical model of quantum mechanics is and *how* to use it. However, the mystery persists for the *why* question, and it is not clear to me whether the solution will be offered by physics, mathematics, or metaphysics. In contrast to quantum mechanics, when scientists are faced with the predictive success of ML modelling, the mystery isn't solely restricted to the question of *why* these ML models work as well as they do. We also need to struggle with the *what*, and the *how* question, and this is what highlights the novelty in the problem of epistemic opacity.

7.4 Epistemic trade-offs in ML modelling in science

In this section, I will review various case studies to understand the nature of epistemic trade-offs that lie at the heart of ML modelling. While scientists might describe the trade-off as a choice between *predictive capabilities* and *explanatory value*, scholars working within XAI and comparing different ML architectures often talk about a similar trade-off between *interpretability* and *accuracy*. Although the very existence of such a trade-off

13. Put simply, the philosophical theme attempts to understand the relationship between representation and reality, like how representations like mathematics and natural language can successfully describe and map onto reality.

has been disputed[14], nevertheless, many scientists do report on the existence of an epistemic trade-off and seem to percieve the issue as being somewhat paradoxical.

The only way that we can understand a trade-off is by comparing multiple cases that prioritize different parts of a trade-off. However, the devil is in the details when experts actually try and compare different models based on their predictive versus explanatory prowess (or between accuracy versus interpretability), because this raises questions on the precise meaning of these fundamental concepts. Marcinkevičs and Vogt (2023) posit that:

> In general, there is no agreement within the ML community on the definition of interpretability and the task of interpretation. (Marcinkevičs and Vogt 2023, 1)

The ML model is to be made interpretable, but interpretable for whom, and to what degree? What does interpretable mean, and can interpretability lead to understanding? Even scientific understanding?

Bell et al. (2022) is another review article which complicates the nature of the trade-off. They talk about an accuracy-explainability trade-off wherein they claim that we cannot even determine explainability of a model based on whether it is a black-box or not! Bell et al. (2022) state that:

> We neither observed a direct trade-off between accuracy and explainability nor found interpretable models to be superior in terms of explainability. It's just not that simple! (Bell et al. 2022, 248)

It is interesting to note that far from resolving differences in opinion about how to approach this puzzle of what to prioritize in a trade-off, the review articles cannot even find consensus on the very nature (or even the existence!) of the trade-off. And even when the reviewers identify a trade-off, they might discover that choosing an explainable model can come at the cost of a very marginal decrease in predictive power:

> The main conclusion of this study is, hence, that although the opaque ensemble models generally obtain higher accuracies than the transparent models, one often has to pay only a limited penalty in terms of predictive performance when choosing an interpretable model instead of an opaque one. (Johansson et al. 2011, 657)

In the case of predictive biopharmaceutical modelling, the outputs of ML models are complemented by information from in vitro or in vivo tests and an explainable

14. Popularly by (Rudin 2019), but also in many of the other case studies that I will presently review.

model is more amenable to such comparisons; simplifying the puzzle for scientists in this particular domain (Johansson et al. 2011).

Reviewers have also pointed out that the post-hoc explanations for the outputs of AI models are not representative of the true causal nature behind the model's result. In their survey on the explainability of ML, Burkart and Huber (2021) conclude thus:

> . . . humans often explain themselves by referring to post-factum coherent stories. Rather than providing two features and an importance of those features with a specified class, the human mind would tend to build a story around those features that explains why it seems obvious that the respective instance belongs to a specific class. . . Thus, the most we can strive for when explaining a model is a sort of human graspable approximation of the decision process. (Burkart and Huber 2021, 301)

Herm et al. (2023) posits that the trade-off between accuracy and interpretability is not well studied. Moreover, whatever preliminary empirical work that has been carried out to address the issue cannot find a smooth function between the epistemic goals of accuracy and interpretability. As the authors note:

> Despite its fundamental importance for human decision-makers, empirical evidence regarding the tradeoff between ML model performance and explainability is scarce. (Herm et al. 2023, 12)

If the evidence for comparing ML models with other ML models is itself "scarce", then the much more important comparison between ML models and conventional models in science must be nearly non-existent!

In Freitas (2019), the author trains a ML model to study other ML models. As useful as this work is for practitioners of ML, I will again reiterate that the major limitation of this work, like all the other scholarship cited above, is that the author can only compare ML models with each other rather than compare ML models with conventional models in science. Although Freitas helpfully highlights the difference between:

1) An interpretable "whitebox" ML model like a decision tree.
2) A "gray box" model like an interpretable ML algorithm deployed in an ensemble.
3) A "black box" model like a neural network.

However, all these models, be it white, gray, or black, are data-driven ML models at their heart. What we need is a comparison between these data-driven ML models with the theory centered models that are conventionally used in science.

Both Freitas (2019) and Herm et al. (2023) assert that scientists can aim to construct more interpretable ML models by making marginal sacrifice in their predictive powers.

However, it is very difficult to figure out a consistent procedure which can quantify the interpretability (let alone explainability) of an ML model. As Freitas (2019) themselves admit:

> . . . this work considered as white box all models using some interpretable knowledge representation, without analyzing the internal details of the models to check if they are really (subjectively) interpretable by users. (Freitas 2019, 65)

Given these complications, it will be challenging to compare data-driven ML models and the more conventional theory centered models in science. Even while comparing among different ML models that vary on the interpretability-complexity grid; it is very hard to determine what should be compared, and how can an expert make this comparison.

The simplest possible survey of this problem would involve training different ML architectures on the exact same scientific data. Such a survey was performed by Wu et al. (2021), however, even in this work, the authors highlight the challenges in comparing complex ML models, like neural networks, with more simpler ones like decision trees. This leads to difficulties in formulating a set procedure that can be used to compare different model architectures in order to determine the trade-off between different epistemic virtues such as the choice between understanding versus model accuracy.

While some reviewers have challenged the existence of a simple trade-off on the performance-interpretability spectrum, others think that such a trade-off exists, and scientists need to be familiar with the nuances of the local context in which they chose to deploy the model (Assis, Véras, and Andrade 2023). As these authors note:

> The evaluated transparent models have better explainability and faster average response times but lower accuracy. In contrast, opaque models have high accuracy, but due to their inherently opaque nature as "black boxes" they lack good explainability and, in all adopted loads, obtained higher average response times than transparent models. (Assis, Véras, and Andrade 2023, 6)

Berenji, Nowaczyk, and Taghiyarrenani (2023) report the conclusions of their work in an interesting way. Like some of the other surveys, they also note how there are "no free lunches" and that any decision regarding the model's architecture will have to deal with trade-offs between interpretability and performance. However, they conclude their survey by stating:

> Therefore, we believe an inherent interpretability versus performance trade-off exists from the data-centric (alternatively to be called representation,

> feature extraction, or preprocessing) perspective. (Berenji, Nowaczyk, and Taghiyarrenani 2023, 52)

This is further evidence for the urgent need to compare between data-centric (or data-driven) models with those of theory-centric (or theory-driven) models. We need to build robust philosophical frameworks to understand and distinguish between "data-centric interpretability", as opposed to the "theoretical interpretability" that is exemplary of conventional scientific practice.

Scientists working on ML modelling themselves see a shift in their disciplinary practice. Karniadakis et al. (2021) conclude their review of the recent methodology of physics-informed machine learning by discussing the possibility of a shift in the fundamental idea of "understanding" in physics. Till now, understanding could be exemplified by the ability to draw interpretations from individual terms in an equation and connect them to observable quantities or other theoretical variables. However, now, scientists can make predictions without having such an understanding. The authors conclude by stating:

> Now, it becomes possible to make accurate predictions without this type of mechanistic understanding—and 'understanding' is something that may be redefined in the process. (Karniadakis et al. 2021, 437)

CHAPTER 8

CONCLUSION

If AI-ML models continue to improve in their epistemic capabilities, they will increasingly be adopted by other scientists across domains. In this thesis, I hope to have motivated the need for further philosophical analysis of how the adoption of these new technologies will impact the scientific method. The widespread adoption of these novel technologies also raises questions about what parts of science are intrinsically valued by humans, and what parts of science can be automated or outsourced to AI. For instance, do we value explanation as an intrinsic goal in science, or will we be content in developing opaque ML models that can make highly accurate predictions but do not provide us with any understanding of *how* they make these predictions?

The concept of *epistemic opacity*—colloquially referred to as the black-box nature of AI-ML programs—is an urgent and philosophically novel concern in contemporary scientific modelling. In the context of ML modelling, I have argued that the epistemic opacity of ML programs necessitates the re-evaluation of fundamental concepts like explanation, prediction, discovery, and so on. Scholars seem to use these concepts in a plurality of ways and the thesis statement posits that these fundamental concepts take on different meanings (compared to their understanding in conventional philosophy of science) in the context of ML modelling in science.

I provide justification for the utility of introducing such a distinction in section 6.4 where I demonstrate the value in acknowleding a gain of epistemic support for the outputs of ML models (prediction$_1$) when they cohere with the results of the reworked theory (prediction$_2$). Similarly, I highlight the utility in distinguishing between:

1) Discovery$_{E1}$, which refers to the creation of a ML model that outperforms the predictions of extant theory. The epistemic value of this discovery lies in its improved predictive accuracy (prediction$_1$).
2) Discovery$_{E2}$, which refers to scientists' success in modifying extant theory in order to reproduce the ML model's predictive capabilities. In the case study of

DeVries et al. (2018), discovery$_{E2}$ comprised the additional theoretical insight into understanding the significance of particular geophysical parameters that led to prediction$_2$.

In chapter 8, I review the epistemic trade-offs in ML modelling in science. One such trade-off is between the predictive accuracy and explanatory value of a ML model. I use my thesis statement[1] as a premise to respond to the confusion among opinions of different scholars regarding the nature (and sometimes, even the existence) of epistemic trade-offs in ML modelling. I posit that one reason behind the confusion is due to lack of clarity in the meaning and application of the epistemic concepts themselves. Although the field of XAI has thoroughly reviewed the concept of explanation, I argue that the concept of prediction and discovery has not been given sufficient attention by philosophers, fueling confusion. I substantiate this set of claims by reviewing various scholarly works opining (often on the basis of empirical results) on the nature of the epistemic trade-offs in ML modelling.

I further argue that the data-driven predictions of opaque ML models are conceptually different from the theoretically centered predictions of conventional science. Working with opaque ML models limits the modes of justification that can be used to support the predictions of ML models. This is one way in which the predictions of ML models are different from the predictions of conventional science. In chapter 8, I demonstrate how the modes of justification for ML discoveries and predictions are different (and limited) compared to the modes of justification for discoveries and predictions of conventional science.

I conclude by motivating the need to build robust philosophical frameworks in order to understand and distinguish between the novel data-driven models (like ML models) versus theoretically centered models that are exemplary of conventional scientific practice.

1. My thesis statement posits that the fundamental concepts of explanation, prediction, and discovery take on different meanings (compared to their understanding in conventional philosophy of science) in the context of ML modelling in science.

BIBLIOGRAPHY

Alvarado, Ramón. 2021. "Explaining Epistemic Opacity." Preprint. https://philsci-archive.pitt.edu/19384/.

Andreas, Holger. 2021. "Theoretical Terms in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University.

Andrews, Mel. 2024a. "The Devil in the Data: Machine Learning & the Theory-Free Ideal."

———. 2024b. "The Immortal Science of ML: Machine Learning & the Theory-Free Ideal."

Assis, André, Douglas Véras, and Ermeson Andrade. 2023. "Explainable Artificial Intelligence - An Analysis of the Trade-offs Between Performance and Explainability." In *2023 IEEE Latin American Conference on Computational Intelligence (LA-CCI)*, 1–6. https://doi.org/10.1109/LA-CCI58595.2023.10409462.

Barberousse, Anouk. 2018. "Philosophy of Physics." In *The Philosophy of Science: A Companion*, 405–29. Oxford Studies in Philosophy of Science. New York (N.Y.): Oxford university press.

Barnes, Eric Christian. 2022. "Prediction Versus Accommodation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Winter 2022. Metaphysics Research Lab, Stanford University.

Beisbart, Claus. 2021. "Opacity Thought Through: On the Intransparency of Computer Simulations." *Synthese* 199 (3): 11643–66. https://doi.org/10.1007/s11229-021-03305-2.

———. 2023. "Computer Simulations." In *Philosophy*. Oxford University Press. https://doi.org/10.1093/obo/9780195396577-0438.

Beisbart, Claus, and Tim Räz. 2022. "Philosophy of Science at Sea: Clarifying the Interpretability of Machine Learning." *Philosophy Compass* 17 (6): e12830. https://doi.org/10.1111/phc3.12830.

Bell, Andrew, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy." In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 248–66. FAccT '22. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533090.

Berenji, Amirhossein, Sławomir Nowaczyk, and Zahra Taghiyarrenani. 2023. "Data-Centric Perspective on Explainability Versus Performance Trade-Off." In *Advances in Intelligent Data Analysis XXI*, edited by Bruno Crémilleux, Sibylle Hess, and Siegfried Nijssen, 42–54. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-30047-9_4.

Bertolaso, Marta, and Fabio Sterpetti, eds. 2020. *A Critical Reflection on Automated Science*. 2020th ed. Human Perspectives in Health Sciences and Technology. Cham, Switzerland: Springer Nature.

Bhatta, Varun. 2023. "Can ChatGPT Be the Author of a Research Paper? – The Wire Science." https://science.thewire.in/economy/tech/can-chatgpt-be-the-author-of-a-research-paper/.

Biever, Celeste. 2023. "ChatGPT Broke the Turing Test — the Race Is on for New Ways to Assess AI." *Nature* 619 (7971): 686–89. https://doi.org/10.1038/d41586-023-02361-7.

Boesch, Brandon. 2024. "Review of Ann Johnson and Johannes Lenhard's *Cultures of Prediction: How Engineering and Science Evolve with Mathematical Tools* - Ann Johnson and Johannes Lenhard , Cultures of Prediction: How Engineering and Science Evolve with Mathematical Tools. Cambridge, MA: MIT Press (2024), 274 Pp. $45.00 (Paperback)." *Philosophy of Science*, November, 1–4. https://doi.org/10.1017/psa.2024.55.

Boge, F. J., P. Grünke, and R. Hillerbrand. 2022. "Minds and Machines Special Issue: Machine Learning: Prediction Without Explanation?" *Minds and Machines* 32 (1): 1–9. https://doi.org/10.1007/s11023-022-09597-8.

Boge, Florian J., and Michael Poznic. 2021. "Machine Learning and the Future of Scientific Explanation." *Journal for General Philosophy of Science* 52 (1): 171–76. https://doi.org/10.1007/s10838-020-09537-z.

Boyd, Nora Mills, and James Bogen. 2021. "Theory and Observation in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University.

Bringsjord, Selmer, and Naveen Sundar Govindarajulu. 2024. "Artificial Intelligence." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2024. Metaphysics Research Lab, Stanford University.

Burkart, Nadia, and Marco F. Huber. 2021. "A Survey on the Explainability of Supervised Machine Learning." *Journal of Artificial Intelligence Research* 70 (January):245–

317. https://doi.org/10.1613/jair.1.12228.

Burkholder, Leslie. 2000. "Computing." In *A Companion to the Philosophy of Science*, edited by W. Newton-Smith, 44–52. Blackwell Companions to Philosophy 18. Malden, Mass: Blackwell Publishers.

Callaway, Ewen. 2024. "Chemistry Nobel Goes to Developers of AlphaFold AI That Predicts Protein Structures." *Nature* 634 (8034): 525–26. https://doi.org/10.1038/d41586-024-03214-7.

Chow, Sheldon J. 2015. "Many Meanings of 'Heuristic'." *The British Journal for the Philosophy of Science* 66 (4): 977–1016. https://www.jstor.org/stable/24562967.

Currie, Adrian. 2015. "Philosophy of Science and the Curse of the Case Study." In *The Palgrave Handbook of Philosophical Methods*, edited by Chris Daly, 553–72. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137344557_22.

Daston, Lorraine. 2008. "On Scientific Observation." *Isis* 99 (1): 97–110. https://doi.org/10.1086/587535.

DeVries, Phoebe M. R., Fernanda Viégas, Martin Wattenberg, and Brendan J. Meade. 2018. "Deep Learning of Aftershock Patterns Following Large Earthquakes." *Nature* 560 (7720): 632–34. https://doi.org/10.1038/s41586-018-0438-y.

Duede, Eamon. 2023. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90 (5): 1089–99. https://doi.org/10.1017/psa.2023.8.

Durán, Juan M. 2018. *Computer Simulations in Science and Engineering: Concepts - Practices - Perspectives*. The Frontiers Collection. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-90882-3.

Durán, Juan Manuel, and Karin Rolanda Jongsma. 2021. "Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI." *Journal of Medical Ethics* 47 (5): 329–35. https://doi.org/10.1136/medethics-2020-106820.

Durán, Juan M., and Nico Formanek. 2018. "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28 (4): 645–66. https://doi.org/10.1007/s11023-018-9481-6.

Facchini, Alessandro, and Alberto Termine. 2022. "Towards a Taxonomy for the Opacity of AI Systems." In *Philosophy and Theory of Artificial Intelligence 2021*, edited by Vincent C. Müller, 63:73–89. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-09153-7_7.

Fahim, Syeda Maham. 2024. "What Is a Floating-Point Arithmetic Problem?" *freeCodeCamp.org*. https://www.freecodecamp.org/news/what-is-a-floating-point-arithmetic-problem/.

Fleisher, Will. 2022. "Understanding, Idealization, and Explainable AI." *Episteme* 19 (4): 534–60. https://doi.org/10.1017/epi.2022.39.

Forster, Malcolm. 2014. "Prediction." In *The Routledge Companion to Philosophy of Science*,

edited by Martin Curd and Stathis Psillos, Second Edition, 449–57. Routledge Philosophy Companions. London ; New York: Routledge, Taylor & Francis Group.

Freiesleben, Timo, Gunnar König, Christoph Molnar, and Álvaro Tejero-Cantero. 2024. "Scientific Inference with Interpretable Machine Learning: Analyzing Models to Learn about Real-World Phenomena." *Minds Mach. (Dordr.)* 34 (3).

Freitas, Alex A. 2019. "Automated Machine Learning for Studying the Trade-Off Between Predictive Accuracy and Interpretability." In *Machine Learning and Knowledge Extraction*, edited by Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, 48–66. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-29726-8_4.

Frigg, Roman, and Julian Reiss. 2009. "The Philosophy of Simulation: Hot New Issues or Same Old Stew?" *Synthese* 169 (3): 593–613. https://doi.org/10.1007/s11229-008-9438-z.

Gadye, Levi. 2019. "Training Computers to Think More Like Scientists." *UCSF School of Pharmacy*. https://pharmacy.ucsf.edu/news/2019/04/training-computers-think-more-scientists.

Godfrey-Smith, Peter. 2003. *Theory and Reality*. Science & Its Conceptual Foundations. Chicago, IL: University of Chicago Press.

Herm, Lukas-Valentin, Kai Heinrich, Jonas Wanner, and Christian Janiesch. 2023. "Stop Ordering Machine Learning Algorithms by Their Explainability! A User-Centered Investigation of Performance and Explainability." *International Journal of Information Management* 69 (April):102538. https://doi.org/10.1016/j.ijinfomgt.2022.102538.

Humphreys, Paul. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. New York: Oxford University Press.

———. 2009. "The Philosophical Novelty of Computer Simulation Methods." *Synthese* 169 (3): 615–26. https://doi.org/10.1007/s11229-008-9435-2.

Johansson, Ulf, Cecilia Sönströd, Ulf Norinder, and Henrik Boström. 2011. "Trade-Off Between Accuracy and Interpretability for Predictive in Silico Modeling." *Future Medicinal Chemistry* 3 (6): 647–63. https://doi.org/10.4155/fmc.11.23.

Johnson, Ann, and Johannes Lenhard. 2024. *Cultures of Prediction: How Engineering and Science Evolve with Mathematical Tools*. Engineering Studies. Cambridge, Massachusetts: The MIT Press.

Karniadakis, George Em, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. "Physics-Informed Machine Learning." *Nature Reviews Physics* 3 (6): 422–40. https://doi.org/10.1038/s42254-021-00314-5.

Kleinberg, Jon, and Sendhil Mullainathan. 2015. "We Built Them, but We Don't Understand Them." In *What to Think about Machines That Think*, edited by John Brockman et al., 68–70. Edge Question Series. New York: Harper Perennial.

Laudan, Larry. 1981. "Why Was the Logic of Discovery Abandoned?" In *Science and*

Hypothesis: Historical Essays on Scientific Methodology, edited by Larry Laudan, 181–91. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-7288-0_11.

Leonelli, Sabina. 2020. "Scientific Research and Big Data." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University.

Marcinkevičs, Ričards, and Julia E. Vogt. 2023. "Interpretability and Explainability: A Machine Learning Zoo Mini-tour." arXiv. https://doi.org/10.48550/arXiv.2012.01805.

Mattioli, Martina, Antonio Emanuele Cinà, and Marcello Pelillo. 2024. "Understanding XAI Through the Philosopher's Lens: A Historical Perspective." arXiv. https://doi.org/10.48550/arXiv.2407.18782.

Messeri, Lisa, and M. J. Crockett. 2024. "Artificial Intelligence and Illusions of Understanding in Scientific Research." *Nature* 627 (8002): 49–58. https://doi.org/10.1038/s41586-024-07146-0.

Mitchell, Melanie. 2025. "LLMs and World Models, Part 1." Substack Newsletter. *AI: A Guide for Thinking Humans*.

Nersessian, Nancy J. 2022. *Interdisciplinarity in the Making: Models and Methods in Frontier Science*. Cambridge, MA: MIT.

Newton-Smith, W. 2000. "Underdetermination of Theory by Data." In *A Companion to the Philosophy of Science*, edited by W. Newton-Smith, 532–36. Blackwell Companions to Philosophy 18. Malden, Mass: Blackwell Publishers.

Nickles, Thomas. 2000. "Discovery." In *A Companion to the Philosophy of Science*, edited by W. Newton-Smith, 85–96. Blackwell Companions to Philosophy 18. Malden, Mass: Blackwell Publishers.

Parker, Wendy S. 2014. "Computer Simulations." In *The Routledge Companion to Philosophy of Science*, edited by Martin Curd and Stathis Psillos, Second Edition, 135–45. Routledge Philosophy Companions. London ; New York: Routledge, Taylor & Francis Group.

Pradeu, Thomas, Maël Lemoine, Mahdi Khelfaoui, and Yves Gingras. 2024. "Philosophy in Science: Can Philosophers of Science Permeate Through Science and Produce Scientific Knowledge?" 75 (2): 375–416.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15.

San Pedro, Iñaki. 2020. "Degrees of Epistemic Opacity." Preprint. https://philsci-archive.pitt.edu/18525/.

Sarukkai, Sundar. 2005. "Revisiting the 'Unreasonable Effectiveness' of Mathematics." *Current Science* 88 (3): 415–23. https://www.jstor.org/stable/24110208.

———. 2012. *What Is Science?* First edition. Popular Readers' Series. New Delhi:

National Book Trust, India.

———. 2015. "To Question and Not to Question: That Is the Answer." In *Public Intellectual in India*. New Delhi, India: Rupa Publications India Pvt. Ltd.

———. 2023. "Philosophy and Method." In *Mapping Scientific Method: Disciplinary Narrations*, edited by Gita Chadha and Renny Thomas, 85–103. Science and Technology Studies. London New York: Routledge.

Schickore, Jutta. 2022. "Scientific Discovery." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Winter 2022. Metaphysics Research Lab, Stanford University.

Srećković, Sanja, Andrea Berber, and Nenad Filipović. 2022. "The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation." *Minds and Machines* 32 (1): 159–83. https://doi.org/10.1007/s11023-021-09575-6.

Strogatz, Steven. 2007. "The End of Insight." In *What Is Your Dangerous Idea?: Today's Leading Thinkers on the Unthinkable*, 130–31. Pymble, NSW: HarperCollins e-books.

———. 2024. "How Is AI Changing the Science of Prediction?" *Quanta Magazine*. https://www.quantamagazine.org/how-is-ai-changing-the-science-of-prediction-20241107/.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73 (1): 109–33. https://doi.org/10.1093/bjps/axz035.

———. 2023. "Do Machine Learning Models Represent Their Targets?" *Philosophy of Science*, October, 1–11. https://doi.org/10.1017/psa.2023.151.

Symons, John, and Ramón Alvarado. 2019. "Epistemic Entitlements and the Practice of Computer Simulation." *Minds and Machines* 29 (1): 37–60. https://doi.org/10.1007/s11023-018-9487-0.

Tal, Eran. 2011. "From Data to Phenomena and Back Again: Computer-Simulated Signatures." *Synthese* 182 (1): 117–29. https://doi.org/10.1007/s11229-009-9612-y.

Turney, Drew. 2024. "GPT-4 Has Passed the Turing Test, Researchers Claim." *Livescience.com*. https://www.livescience.com/technology/artificial-intelligence/gpt-4-has-passed-the-turing-test-researchers-claim.

Van Noorden, Richard, and Jeffrey M. Perkel. 2023. "AI and Science: What 1,600 Researchers Think." *Nature* 621 (7980): 672–75. https://doi.org/10.1038/d41586-023-02980-0.

Vigen, Tyler. n.d. "Bachelor's Degrees Awarded in Engineering Correlates with Electricity Generation in Cambodia (r=0.997)." https://www.tylervigen.com/spurious/correlation/2716_bachelors-degrees-awarded-in-engineering_correlates-with_electricity-generation-in-cambodia. Accessed March 5, 2025.

Wigner, Eugene P. 1960. "The Unreasonable Effectiveness of Mathematics in the Natural Sciences." *Communications on Pure and Applied Mathematics* 13 (1): 1–14. https:

//doi.org/10.1002/cpa.3160130102.

Winsberg, Eric. 2022. "Computer Simulations in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Winter 2022. Metaphysics Research Lab, Stanford University.

Winsberg, Eric B. 2010. *Science in the Age of Computer Simulation*. Chicago: University of Chicago press.

Woodward, James. 2014. "Explanation." In *The Routledge Companion to Philosophy of Science*, edited by Martin Curd and Stathis Psillos, Second Edition, 203–13. Routledge Philosophy Companions. London ; New York: Routledge, Taylor & Francis Group.

Woodward, James, and Lauren Ross. 2021. "Scientific Explanation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2021. Metaphysics Research Lab, Stanford University.

Wu, Leihong, Ruili Huang, Igor V. Tetko, Zhonghua Xia, Joshua Xu, and Weida Tong. 2021. "Trade-Off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets." *Chemical Research in Toxicology*, January. https://doi.org/10.1021/acs.chemrestox.0c00373.