Università degli Studi di Urbino Carlo Bo

Department of Pure and Applied Sciences

Ph.D. PROGRAMME IN: Research Methods in Science and Technology

CYCLE XXXVI

Scientific Understanding, Representations and Models

ACADEMIC DISCIPLINE: M-FIL/02 Logic and Philosophy of Science

Coordinator: Prof. Alessandro Bogliolo

Supervisor: Prof. Vincenzo Fano

Ph.D. student : Giovanni Galli

ACADEMIC YEAR 2022/2023

Contents

Chapter 3: Schurz and Lambert on Scientific Understanding and the Received View

Chapter 4: The Contextual Theory of Scientific Understanding

Chapter 5: Khalifa's Account of Scientific Understanding

Chapter 6: Scientific Understanding and the Explanatory Integration in Cognitive Sciences

Chapter 7: Scientific Representation and Deep-Learning Models: the Case of AlphaFold

Chapter 8: Scientific Understanding and Language Models

Chapter 9: Descriptive and Explanatory Scientific Understanding: Towards a Pluralism of Understanding

Chapter 10: Scientific Understanding and Scientific Realism

Acknowledgements

There are a lot of ideas and notions discussed in this Ph.D. thesis about scientific understanding, representations and models that have been and, in some cases, still form distinctive cornerstones in the Philosophy of Science. Suppose the thesis succeeds in offering an original path towards the relatively recent notion of scientific understanding. In that case, the merit goes to all the friends, Professors, Ph.D. students, and scholars that I gratefully thank for all the discussions and the philosophical acumen they share with passion and integer attention. First, I want to thank the *maestri* I have been lucky to work with in the ReMeST Ph.D. Program at the University of Urbino Carlo Bo, Professor Vincenzo Fano, bracing and active supervisor, and Professor Mario Alai, former supervisor, sharp and passionate, recently retired. I am also grateful to the Ph.D. students of

the ReMeST program, along with Professor Alessandro Bogliolo and Professor Andrea Viceré, for their passionate discussions. Thanks to the mentor Pierluigi Graziani for his endless willingness to improve the ideas and the results of the research. Moreover, I would like to thank Adriano Angelucci, Gabriele Ferretti, Giovanni Macchia, Professor Gino Tarozzi, Stefano Calboli, Davide Pietrini, Mirko Tagliaferri, Professor Alessandro Aldini and the Synergia research group for the invaluable work done together. Thanks also to Professors Raffaele Mascella and Davide Fazio for the inspiring discussion we had at the University of Teramo, still ongoing.

These have been unique and demanding years, I would not have got here without the determined love of Beatrice and all my family: Giulietta, Paolo, Marianna, Dina, Maria, Antonella, Catia and Enrico, Stefano, Giada, Linda, Davide, Maura and Celeste. To all of you, I am deeply thankful. Of course, eventually, the responsibility for all the flaws in this work is mine only.

Introduction

1. Introducing Scientific Understanding

One of the distinctive features of human beings is the desire to understand. Understanding each other, the world, the universe and the phenomena, seems to be very important for us. What is peculiar of our species is that we have developed through hundred of years a specific form of knowledge gathered through the work of research communities that we call science. It is common ground to distinguish different kinds of knowledge in contemporary world, one of which is specific of sciences[1]. We can access to this knowledge thanks to

[1] The plural is used to stress the rich diversity of the scientific knowledge and the many fields and subfields of scientific research, without the implicit deflationary assumption about the absence of a scientific method (Cartwright, Hardie, Montuschi, Soleiman and Thresher, 2022; Boniolo, 1999). This issue is not debated here, but we will see that the debate about scientific understanding is fostered by the aim of unification.

books, articles, lectures, seminars and all the intellectual activities dedicated to the research community. But what happens exactly when scientists overcome the boundaries of what has already been known and open up a new scenario about an (earlier) unknown phenomenon? In other words, before relativity theory and quantum mechanics, how much did we know about the universe? It is easy to answer, we knew less than what we know now. How is it possible to foster our knowledge of something unknown before? Many philosophers of science in the last years started wondering whether what happens in scientific research would be better appreciated, if we don't think about scientific knowledge alone, but also do ask ourselves what it takes to *understand* phenomena in scientific terms. What did we understand about the universe before relativity theory (RT)? And what before quantum mechanics (QM)? Before RT and QM we did not understand the universe as we do now. So, scientific knowledge and scientific understanding seem to be tidily related and the aim of this thesis is to study what is scientific understanding.

2. Motivation and Desiderata

Why understanding is important? The epistemological research about understanding and then the extension of it in the Philosophy of Science about scientific understanding springs from three main reasons, developed on two distinguished conceptual levels: an

axiological point of view and a pragmatic one. The main three reasons are:

1. Understanding seems a central good that we try to realize when we think about the world and the cosmos.
2. Understanding is a central epistemic goal of the sciences.
3. An epistemology that does not restrict itself to knowledge, but takes into account understanding too can do better to accommodate our intuitions about our epistemic achievements in a broad sense (Greco 2010, pp. 7-8).

We can distinguish an axiological level and a pragmatic stance from these three main reasons. The axiological view about understanding states that it is the core of the epistemology research since its origins. After all, humans are animals provided with the ability to understand, and since the famous opening of Aristotle's *Mataphysics*, all the humans by nature tend to understand (*eidénai*)[2]. The Greek verb has been traditionally translated as "knowing", but to a closer view it can be conceived as "understanding", as a more general, dynamic and broad form of knowledge. Several scholars of Ancient philosophy claim that *episteme* has more in common whit what we call "understanding", in fact, according to Annas (1981, ch. 10), for Plato, a person which has *episteme* has a systematic understanding of things, as

[2] See, Grimm (2021).

distinguished to possessing of various true propositions. Similarly, Lear claims that for Aristotle, «to have *episteme* one must not only know a thing, one must also grasp its cause or explanation. This is to understand it: to know in a deep sense what it is and how it has come to be» (Lear, 1988: 6). In this sense, understanding appears more valuable than knowledge, eve if the contemporary scholars ascribe understanding also to low-grade cases, in which we – understanders – may not grasp relevant relations between the elements objects of understanding. This is the famous passage of Lear's reading of the opening sentence of Aristotle's *Metaphysics*:

But if philosophy is the ultimate goal of our original innate desire, perhaps we have to re-think what that desire is. We are not satisfied to know, for example, that the heavens move in such a way; nor will we be satisfied to know a vast array of such facts about the phenomena. We want to know why the heavens move that way, why the phenomena are as they are. We are after more than knowledge, we are after understanding. Aristotle was, I believe, aware of this. Although 'to know' is an adequate translation of the Greek "eidenai", Aristotle used this term generically to cover various species of knowing. One of the species is "epistasthai" (literally, to be in a state of having episteme) which has often been translated as 'to know' or 'to have scientific knowledge', but which ought to be translated as 'to understand'. For Aristotle says that we have episteme of a thing when we know its cause. To have episteme one must not only know a thing, one must also grasp its cause or explanation. This is to understand it: to know in a

deep sense what it is and how it has come to be. Philosophy, says Aristotle, is epistēmē of the truth (Lear 1988: 6).

The pragmatic view about understanding is not to be confused with a specific account of understanding, but has to do with the pragmatic turn in Philosophy of Science, according to which the philosophical analysis and study have to apply to real-cases scientist work on in specific research communities, with peculiar methodologies and research tools. In this line of thinking, "understanding" is a notion multifaced that can be useful to develop the philosophical work about science, focusing on one hand on the scientific activities, and studying the conceptual assumptions and frameworks implied and implemented in the scientist's work. The notion of understanding in this sense can be conceived as a bridge between the realist and antirealist stances about science, and the understanding-reflection can be an alternative solution to instrumentalist and realist tensions.

Moreover, a declination of this last pragmatic point, which inherits the conception of the notion of understanding from the XIX Century, aims at posing understanding as a common field in which sciences and humanities can find and foster a strong cultural alliance[3]. Understanding as a bridge for the alliance between humanities and natural sciences is between the lines of Rovelli (2019), which declares to believe, as scholar, in the

[3] Alliance which is one of the focal point of also Fano's research along the years.

dialogue between the disciplines in the humanities (often seen as suspected from the scientists' point of view) and in the natural sciences:

Le discipline umanistiche sono talvolta viste con sospetto nel mondo delle scienze naturali. Viene loro rimproverata mancanza di chiarezza e dell'oggettività permesse alla matematica e dalle conferme sperimentali. Io appartengo a quella parte di scienziati che ritiene invece che il dialogo tra le scienze naturali e discipline umanistiche sia sempre stato e sia ancora essenziale per entrambi i campi del sapere, e per l'avanzamento della nostra *comprensione*[4] del mondo (Rovelli, 2019: 3).

The dialogue between natural sciences and humanities can be settled, and the philosophical research as an ally to both natural sciences and humanities has the duty – in a deontological sense – to study the cultural intersection of them, which comes through the understanding. If we understand how the scientist work, if we understand scientific understanding, we can also be helpful to strengthen this awaited alliance. Of course there are distinctive kinds of understanding. The one we talk about in the scientific

[4] Emphasis is mine. Rovelli uses the Italian word "comprensione", understanding, to point out that an alliance between sciences and humanities it is possible, and it is thanks to this alliance that is possible the progress of our understanding of the world. Here the translation of Rovelli's lines: «The humanities are sometimes viewed with suspicion in the world of natural sciences. They are reproached for lacking the clarity and objectivity afforded by mathematics and experimental confirmations. I belong to that group of scientists who instead believe that the dialogue between the natural sciences and humanities has always been and still is essential for both fields of knowledge, and for the advancement of our understanding of the world».

areas can have different features than the one we talks about in social epistemology, i.e. the understanding achieved thanks the scientific communication and journalism. While many scholars are working on this issue, namely the understanding in social epistemology, in this pages the focus will be instead on scientific understanding.

Since the study of scientific understanding as a specific area of inquiry has a relevant young age, the first monography[5] on the subject was published in 2009, there are still many open problems to be addressed. There is a common tendence to disentangle the discourse about scientific understanding from the realist framework about the relevance of truth in the scientific activities. Given that the debate about scientific understanding addresses mainly on theories and models, the discussion on scientific representations in the scientific understanding debates is wide. The aim of this research is to add another piece to the puzzle of understanding, trying to make it less untidy with the aid of cases taken from contemporary example of scientific research with technologically advanced theoretical and computational tools. In this pages the principal items of scientific understanding are representations and in particular models.

We will see that some more reflection is needed in order to distend the tension between antirealistic assumptions and realistic ideas in this debate. Even if most of the scholars want to rule out

[5] See De Regt, Leonelli and Eigner, (2009).

the truth requirement for scientific understanding, I still claim that it cannot be undone, specifically also for particular models, such as deep-learning models.

The main purpose of this research is to contribute to the discussion about scientific understanding, making room for the awareness in the philosophical community that scientific understanding is a notion worth to be scrutinized, given the intrinsic features we all human beings shared: «Nature is what it is, while we as biological organisms try to understand as best we can what we experience» (Danielsson, 2023: 12). To try to understand as best we can it takes time; it not by chance that we knew the timing and the place of the total eclipse of April 8, 2024, after thousands of years of observations, speculations, data collection and scientific study. We can predict the occurrence of a solar eclipse as a result of «an intricate overlap of orbital cycles that took millennia to *understand* [emphasis is mine]» (Sokol, 2024). It is worth dedicating our attention to the notion of scientific understanding, as a fundamental part of our lives of biological organisms that try to understand scientifically – as best we can – what we experience (Danielsson, 2023). Danielsson's assertion underscores the profound significance of scientific understanding in the human experience. As biological organisms driven by innate curiosity, we are compelled to make sense of the world around us through the lens of scientific inquiry. By striving to understand the intricate workings of the universe to the best of

our ability, we engage in a continuous quest for knowledge and information. Indeed, scientific understanding serves as a cornerstone of human cognition, enabling us to navigate the complexities of our existence with greater clarity and insight. Therefore, dedicating attention to the notion of scientific understanding is not only essential for advancing our understanding of the natural world but also for enriching our own lives as human beings.

3. Structure of the Thesis

The focus of the thesis is scientific understanding an I will discuss in particular the characteristics of two kinds of scientific understanding: explanatory and descriptive scientific understanding. The descriptive understanding will be discussed using examples of deep-learning models in bioinformatics and computational linguistics. The topics of scientific understanding, representation and models in science are overly broad and complex. It is important to note, for the readers' clarity, that I will focus on a particular intersection of these notions, which it is helpful to deal with some novel issues raising in the philosophy of science and in the philosophy of AI in the last years. So, this is not a thesis on general issues about scientific understanding, or about the key problems of representation and models in science, but it is a work about the

scientific understanding and its features we, as inquirers, can gain using AI systems and tools throughout the scientific research.

From this perspective, the structure of the thesis is meticulously organized to delve deeply into the multifaceted nature of scientific understanding, with a particular emphasis on elucidating the characteristics of two distinct types of scientific understanding: explanatory and descriptive. This comprehensive exploration unfolds across a series of thoughtfully crafted chapters, each contributing to a nuanced understanding of the subject matter.

Beginning with Chapter 1, the thesis lays a solid foundation by scrutinizing the intricacies of scientific explanations, setting the stage for subsequent discussions. Chapter 2 delves into the philosophical concept of Verstehen, providing a theoretical framework for understanding the cognitive processes involved in grasping scientific concepts.

Chapters 3, 4 and 5 offer critical analyses of the perspectives of notable scholars such as Schurz, Lambert, de Regt, and Khalifa, respectively. These chapters serve to illuminate the diverse array of viewpoints within the realm of scientific understanding, enriching the discourse with a multiplicity of insights and perspectives.

In Chapter 6, the focus shifts towards practical applications, as the thesis explores the concept of explanatory integration within the field of cognitive science. This chapter examines how different explanatory models can be integrated to enhance our understanding of complex cognitive phenomena.

Chapters 7 and 8 pivot towards empirical examples, with discussions centred on the AlphaFold models in bioinformatics and language models in computational linguistics, respectively. These case studies offer tangible illustrations of how scientific understanding is manifested in real-world contexts, shedding light on the practical implications of theoretical concepts.

The thesis culminates in Chapters 9 and 10, where the distinctions between descriptive and explanatory scientific understanding are thoroughly explored. Through in-depth analyses and critical reflections, these chapters provide a comprehensive overview of the nuanced relationship between understanding and representation in the scientific domain.

Overall, the structured approach of the thesis facilitates a systematic exploration of scientific understanding, guiding the reader through a rich tapestry of theoretical insights, empirical examples, and philosophical reflections. By engaging with a diverse array of perspectives and disciplines, the thesis offers an extensive examination of scientific understanding in contemporary contexts of scientific inquiry.

Chapter 1:
Explanation and Understanding

1.1. Introduction

On September 14 2015, at 9:50:45 UTC, the Ligo interferometer observed a transient gravitational-wave signal. It was a remarkable detection and a marvellous output of a great experiment that confirmed what Einstein predicted about the fabric of space-time. It took a hundred years to verify a theoretical claim in the most advanced physics of the first half of the twentieth century. In the paper published a year after the observation, the researchers (Abbott et al., 2016: 1) wrote as follows:

> In 1916, the year after the final formulation of the field equations of general relativity, Albert Einstein predicted the existence of gravitational waves. He found that the linearized weak-field

> equations had wave solutions: transverse waves of spatial strain that travel at the speed of light, generated by time variations of the mass quadrupole moment of the source. Einstein *understood*[6] that gravitational-wave amplitudes would be remarkably small; moreover, until the Chapel Hill conference in 1957 there was significant debate about the physical reality of gravitational waves.

Abbott et al. (2016: 1) in their opening lines of the famous paper, that generated a huge debate about gravitational waves and the urgency to stabilize the international scientific program regarding the development of new interferometers on earth and in space, quote explicitly Einstein's understanding of the existence of gravitational waves and their amplitude.

It would be unsurprising to find a remarkable number of quotations of the verb "to understand" and its derivative substantive "understanding" along the lines of scientific papers, books, and interviews. It is the aim of this thesis to study the specific kind of understanding involved in the scientific activities, namely the scientific understanding. It is the kind of understanding scientists come to during the scientific research.

The experiment leading to this astonishing result has taken many years since the hypothesis of the existence of gravitational waves was advanced by Einstein (1916). This result has given the greatest contribution to the confirmation of the hypothesis and as a proof of the existence of the gravitational waves also assesses a

[6] Emphasis is mine.

higher level of understanding of the phenomenon then the theoretical understanding of the hypothesis. In the first half of the XX century we could have a theoretical understanding of the phenomenon, given the scientific knowledge stated by Einstein regarding the hypothetical existence of gravitational waves – so we had a theoretical knowledge of the hypothesis of their existence – but only thanks to the great long-lasting experimental effort we now have data, proofs and demonstrations of their existence. The jump from the theoretical knowledge and theoretical understanding of the phenomena, through the hard work of Abbott's team and all the researchers involved all around the world in the detection of gravitational waves, is now done.

I think this example suits well to introduce the central theme of this thesis, scientific understanding of phenomena given by experts and scientists working painlessly indeed to understand scientifically the objects they study, their causal relations, the correlations and their functions within the dynamic framework of the systems in which they act. The entanglement between scientific understanding and scientific knowledge is very important and tight. We will see in these pages that some scholars are inclined to decouple the two notions, while others argue that the notions are deeply tied to each other.

Still, understanding the world, the universe, the cosmos and its phenomena is deeply connected to what we are as humans.

Looking at the sky is something humankind has been doing since its origins, with eyes full of wonder:

«Perhaps from that wonder was born the spark that ignited curiosity towards the Sky. Then, probably, in a distant past, something deeper took over, a desire to seek regularity in celestial events, like the rising or setting of the Sun, the cycle of lunar phases, or that of the seasons. Not only out of a thirst for knowledge but also to make that primitive world that appeared random and sometimes dangerous more understandable and reassuring».[7] (Razzano, 2021: 9).

According to Razzano, from the wonder of the cosmos comes the desire to search for the regularities in the events, the phenomena, not only to know, but also to make understandable the primitive world that before the work of scientists appeared casual and dangerous. What I am going to expose in this thesis is an analysis of what is the wonderful endeavour that can make the world and the cosmos understandable, namely scientific understanding. So, in the following pages, the understanding inquiry of scientists will be the focus.

[7] See Razzano (2021: 9) for a clear description (for the Italian public) of the newest discoveries and advancements in contemporary astrophysics. His opening lines show the desire to understand that characterises the humankind: «Forse da quella meraviglia è nata la scintilla che ha acceso la curiosità verso il Cielo. Poi, probabilmente, in un lontano passato subentrò qualcosa di più profondo, un desiderio di cercare una regolarità negli eventi celesti, come il sorgere o il tramontare del Sole, il ciclo delle fasi lunari o quello delle stagioni. Non solo per sete di conoscenza, ma anche per rendere più comprensibile e rassicurante quel mondo primitivo che appariva casuale e a volte pericoloso».

Scientific understanding seems to be then something more than a description of how the things in the world and the universe are. But this seems to be just one of its aims. While on the one hand, sciences aim at describing the phenomena, on the other hand, sciences also strive to explain why phenomena are such and such. A discussion about scientific understanding would be incomplete without further discussion of scientific explanations.

Before we introduce the two notions of understanding and scientific understanding, let me take a detour on scientific explanations that, in the last Century, attracted a lot of philosophical attention thanks to the work of famous scholars such as Hempel, van Fraassen and Salmon.

1.2. Scientific Explanations

Every day of our lives, we have to deal with explanations. Driving on the way to work, my car stops, and I wonder why it does not work anymore. Or, why Adem does not show up at school today? One of the scientists' jobs is to produce explanations, and scientific explanations come in many forms insofar as they are elaborated in different ways and distinct scientific areas.

The first clue of scientific behaviour is a question mark. Indeed, scientific research involves investigating phenomena

through questions. Philosophers of science distinguish different kinds of explanations as answers to why-questions. But not all why-questions demand an explanation, but in the case of scientific questions, gaining explanations is the core of the research activity.

In Chapter 2, the discussion will deal with the conceptual differences between explanations, scientific understanding and *Verstehen* (the German word for understanding). This discussion has been central to Von Wright's book *Explanation and Understanding* (1971), in which he distinguishes two main traditions defining explanations: the Aristotelian stressing on explanations as purpose and intentions, and the Galilean ways taking causal explanations as the cornerstone. After a detailed analysis of causal explanation, intentionality and teleological, and explanation in history and the social sciences, Von Wright concludes that explanations about human actions – the ones that according to Dilthey are to be understood and not explained – explanations which are fundamental in human sciences cannot be reduced to causality. There is a *fil rouge* bonding the debate about explanations, *Verstehen*, and scientific understanding that we have to follow and unravel, in order to propose a prospective study about scientific understanding. From now on, in this chapter, I present the main models of explanation we can find in the philosophy of science of the last century, namely the Deductive-Nomological Model, the Probabilistic and Statistical Relevance

Model, the Causal-Mechanical, the Functional, the Pragmatic model, and the Unification view.

1.2.1 The Deductive-Nomological Model

The Deductive-Nomological model (ND) takes its name from Hempel and Oppenheim's essay (1948) about the explanation of scientific facts and laws. This model of explanation, also known as the covering law model, stands as one of the most influential explanation models in the philosophy of science. Developed by Carl Hempel in the mid-20th century, this model seeks to elucidate the structure of scientific explanations by emphasizing the role of laws and logical deduction. At its core, the model posits that scientific explanations consist of two essential components: a set of general laws or regularities and initial conditions or specific facts. According to Hempel, an explanation is achieved when the occurrence of a particular event is logically deduced from these general laws and initial conditions. This deductive process, known as subsumption, entails demonstrating that the event to be explained is a logical consequence of the laws and conditions, thereby "covering" it in a logical sense. The model's name, "deductive-nomological", derives from its reliance on nomological (law-like) principles and deductive reasoning.

Let's use the well-known example Hempel takes from J. Dewey. He wants to explain the fact *E*: after washing a glass in a sink with hot water and soap, we turn it and lay it upside down on a tray. We will see soap bubbles forming between the rim of the glass and the tray surface. The bubbles will increase their shape, get smaller and then disappear. Hempel and Oppenheim develop ND model, to describe the inferential structure of the explanation that from the premises, given by laws and initial conditions, leads to the conclusion which is the fact we want to explain. According to the example above, the ND model will be:

Covering Laws

L_1 laws about gases;

L_2 laws about heat transmission;

L_3 laws about the behaviour of elastic bodies

[…]

Initial Conditions

C_1 – the glass is immersed in hot water;

C_2 – the glass is upside down and lay down on a tray;

C_3 – between the rim of the glass and the tray there is a film of foamy water

Explanandum

E – the formation of soap balls, their expansion and behaviour.

In this scheme, the *explanans* consists of two parts: the general laws (L_1,…L_j), on which the explanation is based, and the initial conditions. The *explanandum* obtains through deductive inference from the *explanans*:

Covering laws	$L_1,\ldots, L_j$
Initial Conditions	$C_1,\ldots, C_k$
	_____
Explanandum	E

Central to Hempel's model is the notion of scientific laws, which are universal generalizations that describe regularities or patterns in nature. These laws are typically formulated as empirical generalizations derived from observation and experimentation. For instance, Newton's laws of motion or Kepler's laws of planetary motion exemplify such universal regularities in physics. In the deductive-nomological model, these laws serve as the fundamental building blocks of scientific explanations, providing the theoretical framework within which specific phenomena are understood.

Complementary to the role of scientific laws are the initial conditions, which represent the specific circumstances or states of affairs relevant to the phenomenon being explained. These initial conditions specify the particular context in which the laws operate and play a crucial role in determining the outcome of the

deductive process. Without accurate initial conditions, even the most robust scientific laws may fail to yield valid explanations. Thus, Hempel emphasizes the importance of meticulous empirical observation and measurement to establish the precise initial conditions necessary for explanatory purposes.

The process of explanation within the deductive- nomological model unfolds through logical deduction. Given a set of general laws and initial conditions, scientists employ deductive reasoning to derive the occurrence of the event or phenomenon. This deductive inference proceeds by subsuming the specific case under the general laws, demonstrating that the event follows logically from the conjunction of the laws and initial conditions. Importantly, this deductive process is characterized by its necessity: if the laws and initial conditions are true, then the event's occurrence must logically follow. Thus, scientific explanations within the deductive-nomological model are claimed to be logically rigorous and deterministic. According to Hempel and Oppenheim, any scientific explanation must satisfy two adequacy conditions: logical and empirical adequacy.

Conditions of logical adequacy:

1. the *explanandum* must be a logical consequence of the *explanans*;
2. the *explanans* must contain at least one covering law;

3. the *explanandum* must be verifiable independently of the *explanans*.

Condition of empirical adequacy:

4. the statements forming the *explanans* must be true.

Critics of the deductive-nomological model have raised several objections and challenges to its adequacy as a scientific explanation theory. One prominent critique concerns the issue of explanation asymmetry. This is the case of the flagpole counterexample proposed by Bromberger (1962), according to which the length of its shadow projected on the ground by the sunlight can be explained using the ND model with some covering laws, such as the geometric optics, and certain initial conditions, as the length of the pole and the position of the sun. In this case, we have an explanation of the length of the shadow that satisfies the adequacy conditions. Anyway, on the base of the ND model, it is not only possible to explain the length of the shadow in virtue of the pole, but also, vice-versa, the length of the pole in virtue of its shadow. The ND model has a problem as far as the asymmetry between the length of the shadow and the length of the pole shows that the adequacy conditions are not sufficient to warrant the soundness of the ND explanation.

While the model appears well-suited to explaining events in terms of deterministic laws, it seems less capable of accounting

for probabilistic or statistical regularities, which are pervasive in many scientific domains. Moreover, some critics argue that the model's emphasis on deductive subsumption overlooks the role of causal mechanisms and contextual factors in scientific explanation. In many cases, understanding why an event occurs may require more than just deductive inference from general laws and initial conditions; it may necessitate a deeper grasp of the underlying causal processes at work[8].

In response to these criticisms, proponents of the nomological-deductive model have sought to refine and extend its conceptual framework. One avenue of development involves incorporating probabilistic reasoning into the deductive structure of explanations, thereby accommodating stochastic phenomena within the model. Additionally, efforts have been made to complement deductive subsumption with causal-mechanical explanations that highlight the mechanisms underlying observed regularities. By integrating these elements, advocates of the model aim to enhance its explanatory power and applicability across a broader range of scientific disciplines.

Despite its limitations and ongoing debates, Hempel's deductive-nomological model remains a cornerstone of the philosophy of science, offering valuable insights into the nature of scientific explanation. Its emphasis on the systematic application

[8] This is one of the arguments the proponents of scientific understanding could use to widen the framework of the theories of explanation to account also for understanding in scientific contexts.

of general laws and deductive reasoning has profoundly influenced how we conceptualize and evaluate explanations in diverse fields of inquiry. While the model may not provide a comprehensive account of all scientific explanations, it continues to inspire fruitful dialogue and theoretical developments within the philosophy of science, shaping our understanding of the rationality and methodology of scientific inquiry.

1.2.2. Probabilistic Explanations and Statistical Relevance

Probabilistic explanations in the philosophy of science represent a fascinating intersection of epistemology, ontology, and methodology, offering nuanced insights into the nature of scientific inquiry and the structure of reality itself. At its core, the probabilistic framework acknowledges the inherent uncertainty that permeates scientific endeavours, recognizing that absolute certainty is often elusive in the study of complex systems. Rather than seeking definitive, deterministic explanations, aspiring at the universality of the covering laws as in the DN model, probabilistic approaches embrace the probabilistic nature of phenomena, acknowledging that events unfold within a range of possible outcomes, each with its associated likelihood. This recognition of uncertainty challenges traditional notions of causality, inviting a

deeper exploration of the probabilistic relationships that govern the behaviour of natural and social phenomena[9].

The concept of probability itself is central to the discussion of probabilistic explanations, which serves as a fundamental tool for quantifying uncertainty and reasoning about the likelihood of different outcomes. Within the philosophy of science, probabilities are understood not merely as mathematical abstractions but as epistemic tools that reflect our knowledge and beliefs about the world. Bayesian probability theory, in particular, offers a powerful framework for updating beliefs in light of new evidence, providing a coherent framework for inference and decision-making under uncertainty. By treating probabilities as expressions of subjective degrees of belief, Bayesianism emphasises the role of evidence in shaping our understanding of the world, highlighting the inherently subjective nature of scientific knowledge.

Probabilistic explanations also raise profound ontological questions about the nature of causality and determinism. While classical Newtonian physics operated within a strictly deterministic framework, the advent of quantum mechanics shattered this deterministic worldview, revealing a fundamentally probabilistic nature at the heart of physical reality. Quantum indeterminacy, famously encapsulated in Heisenberg's uncertainty

[9] Rescher (1962) was one of the first scholars to focus on the difference between the ND model and the actual practice of statistical laws in science.

principle, suggests that at the subatomic level, events unfold probabilistically, with outcomes governed by probability distributions rather than definite causal laws. This radical departure from classical determinism challenges traditional notions of causality, prompting philosophers to reconsider the nature of causation within a probabilistic framework. In this light, causation is understood not as a deterministic relation between cause and effect but as a probabilistic tendency for certain events to follow others, contingent upon various factors and contextual conditions.

The embrace of probabilistic explanations in the philosophy of science also has significant implications for scientific methodology and practice. Researchers grapple with inherent uncertainties and complexities that defy simple deterministic explanations in fields ranging from physics to biology to economics. Probabilistic models provide a flexible and robust means of capturing this complexity, allowing scientists to make predictions and infer causal relationships in the face of uncertainty. Statistical inference techniques, such as hypothesis testing and parameter estimation, provide formal methods for evaluating the strength of evidence and making inferences about underlying processes based on observed data. Bayesian methods, in particular, offer a coherent framework for integrating prior knowledge with new evidence, enabling researchers to update their beliefs and make informed decisions in light of uncertainty.

Despite its many virtues, the probabilistic framework poses challenges and raises crucial philosophical questions. One such challenge is the problem of induction, famously articulated by the Scottish philosopher David Hume. Hume argued that the justification for inductive reasoning, which underpins much of scientific inference, ultimately relies on circular reasoning or unjustified assumptions about the uniformity of nature. While probabilistic approaches offer a partial solution to this problem by quantifying degrees of support for hypotheses based on evidence, they do not fully resolve the underlying epistemological uncertainty inherent in induction. This tension between the need for reliable inference and the inherent uncertainty of induction remains a central concern in the philosophy of science.

Another philosophical challenge posed by probabilistic explanations is the interpretation of probability itself. While frequentist interpretations define probability in terms of long-run frequencies of events, Bayesian interpretations treat probability as a subjective degree of belief. These differing interpretations raise questions about the epistemic status of probabilities and the relationship between subjective belief and objective reality. Moreover, the interpretation of probability within quantum mechanics poses further challenges, with competing interpretations such as the Copenhagen interpretation, the many-worlds interpretation, and the pilot-wave theory offering starkly

different accounts of the probabilistic nature of quantum phenomena.

In conclusion, probabilistic explanations represent a rich and multifaceted approach to understanding the nature of scientific inquiry and the structure of reality. By embracing uncertainty and acknowledging the probabilistic nature of phenomena, probabilistic approaches offer a flexible and robust framework for reasoning about complex systems and making predictions in the face of uncertainty. However, the embrace of probabilistic explanations also raises profound philosophical questions about the nature of causality, the justification of induction, and the interpretation of probability itself. As scientists continue to grapple with uncertainty and complexity in their quest to understand the natural world, the philosophical insights offered by probabilistic explanations remain as relevant and thought-provoking as ever.

On the basis of the reflections about Hempel's conditions on probability and the role of relevance in a statistical explanation, Salmon (1965; 1984) came up with an alternative model, named the Statistical Relevance (SR) model (Salmon, 1971). The SR model is based on two observations:

1. not every explanation is an argumentation in which the facts are inferred from certain premises.

2. The explanation is more or less sound, depending on the quality and quantity of relevant information contained in the *explanans*.

The idea of tracking the relevance of the utterances of specific cases comes from the objections that the high probability of covering law is neither a sufficient nor a necessary condition. What is important is that the probability be recorded on a statistically relevant sample in relation to the case to explain. In the example discussed by Salmon (1989: 63), John Jones has an infection caused by strep (streptococcus) bacteria that could be penicillin-resistant or not. The underlying ideas of the example is that to assume covering laws for a set of cases, those have to be statistically relevant. In fact, to explain the case, we should consider the relevant samples of people infected by strep penicillin-resistant, not by strep bacteria in general. To evaluate such statistical relevance, we should take two probabilities, one related to the cases of infected by general strep bacteria, and one to the cases of infected by penicillin-resistant strep. It means that we should divide the statistical sample in different cells. Following Salmon's example, we have:

a) S = reference set of individuals infected by streptococcus bacteria;
b) $P(H/S)$ = the probability of healing for the individuals in S.

We could then partition S through subclasses like (C): the cell of individuals receiving the penicillin; and like ($\neg C$): the cell of individuals that have not been cured with penicillin. So, we have $S = C \cup \neg C$. In particular, a partition of S is obtained when we have a set of subsets not empty of C such that:

1. $S_i \cap S_j = \emptyset$, for every $i \neq j$;
2. ${}^{n}\cup_{i=1} S_i = S$.

It means that the partition is statistically relevant in comparison to the characteristic H, if the probability of H is different for each cell. So, $p(H/S_i) \neq p(H/S_j)$, for every $i \neq j$. In this way, we will have also two probabilities, one *prior* given by $p(H/S)$, and one *posterior* given by $p(H/S_i)$. If we compare them, we will have three cases:

1. If $p(H/S_i)$ is $>$ $p(H/S_j)$, S_i is positively relevant for H;
2. If $p(H/S_i)$ is $<$ $p(H/S_j)$, S_i is negatively relevant for H;
3. If $p(H/S_i)$ is $=$ $p(H/S_j)$, S_i is irrelevant for H.

The example says that John, with strep infection, received penicillin and healed. So, recalling the two classes (C) and ($\neg C$) and the partition of S, we see that if $p(H/C)$ is $>$ $p(H/S)$, then the cell C is positively relevant for H. But, we need to distinguish one more cell, to account for the possibility that some strep could be

penicillin-resistant. Be then the set S partitioned in other two cells (R): the individuals infected by strep penicillin-resistant; and ($\neg R$): the individuals infected by strep non penicillin-resistant. Together with the first partition, we will have four cells:

1. $C_1 = C \wedge R$
2. $C_2 = C \wedge \neg R$
3. $C_3 = \neg C \wedge R$
4. $C_4 = \neg C \wedge \neg R$

In this way, all the individuals with strep infection are partitioned in cells, but overall the partition is not relevant, due to the same probability value of some cells. Indeed, if John is cured with penicillin but has strep penicillin-resistant, or if John is not cured either with penicillin-resistant or non resistant strep, it has the same probability. Only the cell C_2 is statistically relevant.

So, to capture the conditions for a statistically relevant partition, S has to be divided in two cells:

a. $C_1' = C \wedge \neg R$ (individuals recovered with penicillin and infected with strep non penicillin-resistant;
b. C_2' = 1, or 3, or 4 = $(C \wedge R) \vee (\neg C \wedge R) \vee (\neg C \wedge \neg R)$.

The relevance of the partition we want to focus on is given by the relation between the cells a. and b., that shows:

$$p(H/C_1') \neq p(H/C_2')$$

and it means that the partition is relevant and that we can explain the recovering of John with its inclusion in the cell C_1'. The SR model can be described with the relation between the prior probability $P(H/S)$ respect to characteristic H taken into account and the posterior probability $P(H/C_1')$ concerning the partition C_1' of the initial class S, relevant respect to H. If there is a relevant partition respect to H of the initial class, then we have a statement about the cell of the predicative function Hx.

This model, as the other we have seen, has its limits. Salmon recognizes that with an explanation of this sort we could confuse the statistical relevance with the causal relevance. It is a confusion at the base of the fallacy in which a statistical randomness is taken as a relevant causal correlation.

1.2.3. Causal-Mechanical Explanations

The flows of the models we have seen so far point out that the explanation would better be not an inference, but the identification of the cause of a phenomenon (Salmon, 1984). Indeed mechanistic, causal explanations occupy a central position in the philosophy of science, offering a robust framework for understanding the underlying structure and dynamics of natural

phenomena. At their core, mechanistic explanations[10] seek to uncover the causal mechanisms that give rise to observed patterns of behaviour, aiming to elucidate the underlying processes that govern the behaviour of complex systems. Unlike purely descriptive accounts that focus solely on correlational relationships, mechanistic explanations delve deeper, identifying the specific causal pathways and interactions that lead to the emergence of observed phenomena. This emphasis on causality reflects a fundamental commitment to uncovering the underlying principles that govern the behaviour of the natural world, seeking to move beyond mere patterns of association to uncover the deeper, underlying structures that give rise to observed phenomena.

Central to mechanistic explanations is the notion of causality itself, which lies at the heart of scientific inquiry. Causality represents the fundamental relation between cause and effect, capturing the idea that certain events or conditions give rise to others in a predictable and law-like manner. Mechanistic explanations seek to uncover these causal relationships, identifying the specific mechanisms and processes through which one event or condition brings about another. This emphasis on causality distinguishes mechanistic explanations from purely statistical or correlational accounts, which focus solely on patterns of association without delving into the underlying causal

[10] See Salmon (1994) and Dowe (2000).

structures that give rise to these patterns. By uncovering the causal mechanisms that underlie observed phenomena, mechanistic explanations provide a deeper understanding of the natural world, offering insights into the fundamental principles that govern its behaviour.

One of the key strengths of mechanistic explanations lies in their ability to provide a detailed and comprehensive account of the underlying processes that give rise to observed phenomena. By identifying the specific causal mechanisms at work, mechanistic explanations offer a level of explanatory depth that is often lacking in purely descriptive or correlational accounts. Rather than merely identifying patterns of association, mechanistic explanations seek to uncover the underlying processes and interactions that lead to these patterns, providing a more nuanced and detailed understanding of the natural world. This emphasis on explanatory depth enables mechanistic explanations to provide insights into the underlying principles that govern the behaviour of complex systems, shedding light on the fundamental mechanisms that give rise to observed phenomena.

In addition to their explanatory depth, mechanistic explanations also offer a degree of generality and predictive power that is often lacking in purely descriptive accounts. By uncovering the underlying causal mechanisms that govern the behaviour of complex systems, mechanistic explanations provide a basis for making predictions and extrapolating beyond observed data.

Rather than relying solely on empirical observations, mechanistic explanations enable scientists to infer the behaviour of systems under novel conditions, extrapolating from known causal mechanisms to make predictions about how systems will behave in the future. This predictive power stems from the fact that mechanistic explanations are grounded in an understanding of the underlying causal principles that govern the behaviour of the natural world, allowing scientists to generalize beyond specific instances to uncover more general principles that apply across a wide range of contexts.

Despite their many virtues, mechanistic explanations also face a number of challenges and criticisms within the philosophy of science. One of the key challenges lies in the identification and delineation of causal mechanisms themselves. While mechanistic explanations seek to uncover the specific mechanisms and processes that give rise to observed phenomena, the identification of these mechanisms is often fraught with uncertainty and ambiguity. Complex systems often exhibit emergent properties that cannot be reduced to the behaviour of individual components, making it difficult to identify the specific mechanisms that give rise to observed phenomena. Moreover, the boundaries between different causal mechanisms are often blurred, with multiple mechanisms interacting in complex ways to produce observed patterns of behaviour. As a result, the identification and delineation of causal mechanisms represent a significant challenge

for mechanistic explanations, requiring careful empirical investigation and theoretical analysis to uncover the underlying causal structures that govern the behaviour of complex systems.

Another challenge facing mechanistic explanations lies in their ability to account for the role of context and contingency in shaping causal relationships. While mechanistic explanations seek to uncover the underlying causal mechanisms that govern the behaviour of complex systems, they often overlook the role of contextual factors and contingent events in shaping these mechanisms. Complex systems are often sensitive to initial conditions and external influences, leading to nonlinear and unpredictable behaviour that cannot be fully captured by mechanistic explanations alone. Moreover, the behaviour of complex systems is often contingent upon a wide range of contextual factors, such as environmental conditions, historical events, and social interactions, which can shape the operation of causal mechanisms in unpredictable ways. As a result, mechanistic explanations must grapple with the role of context and contingency in shaping causal relationships, recognizing that the behaviour of complex systems is often shaped by a wide range of factors that cannot be fully captured by mechanistic models alone.

In conclusion, mechanistic, causal explanations occupy a central position in the philosophy of science, offering a robust

framework for understanding the underlying structure and dynamics of natural phenomena. By uncovering the underlying causal mechanisms that govern the behaviour of complex systems, mechanistic explanations provide a deeper and more comprehensive understanding of the natural world, offering insights into the fundamental principles that govern its behaviour. However, mechanistic explanations also face a number of challenges and criticisms, including the identification and delineation of causal mechanisms, and the role of context and contingency in shaping causal relationships. Despite these challenges, mechanistic explanations remain a powerful tool for understanding the natural world, offering a detailed and nuanced account of the underlying causal structures that give rise to observed phenomena.

1.2.4. Functional Model of Explanations

On the other side of the dialectics of theory of explanations, there is a different model of explanation that does not consider the specific laws contained in the *explanans*. Instead, it explains fact, laws or phenomena in virtue of their function. Thus, it is the functional model (FM) of explanation[11]. A functional explanation can be the one we use to answer why human beings have lungs,

[11] See Nagel (1961) for a definition of functional model.

asking about the role (or function) of lungs in the human body. In the answer to this question, it is important to describe the role of oxygen in the human metabolism and the role of lungs in the exchange of oxygen in the air and blood. In this kind of explanation it is not interesting to describe the correct anatomy and functioning of lungs, but to explain their role for such biological activities in human body. The pattern and regularity are not the focus here, but the model aims to capture the necessary and sufficient conditions for a biological, physical, social, artefactual or human process.

The functional explanation involves the notions of system S, organization O, environment E, and system equilibrium or disequilibrium of the system. The functional explanation opens up to an innovative way of studying the natural and artefactual phenomena, from the point of view of the systemic approach.

The systemic approach has three important features to consider. The first 1) is that every item of the system can be conceived as a system itself of a lower level, so each system can be a subsystem of a higher level. This means that the systemic approach poses a hierarchy. From an epistemological point of view, the chosen level of the hierarchy to start with should not limit the scientific inquiry, but it must be open to shift and change from level to level in methodology and dynamics. The second 2) concerns the feedback mechanisms[12] the functional analysis establishes. When a

[12] See Wiener (1948) about the feedback loops.

parameter of the system changes, the interaction between each part of the system makes the whole system varies. The 3) feature is that the functional explanation does not concern all the possible relations of cause and effect between the properties of the system, but it focus on a limited set of relevant parameters.

So, in the functional explanation we would not see a pattern like the cause-effect relation, but we will have the state A related to the state B, that can change and influence in turn the state A (it is the case of the teleological relation between the radiator, the temperature in the room and the thermostat). This is one of the explanation model that fit better to many different scientific research areas different from natural sciences (Boniolo, 2003: 132), and which is flexible and does not rely on the application of covering laws.

1.2.5. Explanations as Unification

Explanations as unification stand as a cornerstone in the philosophy of science, embodying the quest to reveal underlying patterns and connections that unify seemingly disparate phenomena[13]. At its essence, unification in science seeks to integrate diverse empirical observations and theoretical frameworks under a unified explanatory framework, thereby

[13] See Fano (2005: 23-26); Friedman (1974) and Salmon (1989).

reducing complexity and providing a coherent understanding of the natural world. This pursuit of unity reflects a fundamental aspiration to uncover the underlying simplicity and elegance that govern the complexity of nature, revealing deeper patterns and regularities that transcend superficial diversity. Unification manifests in various forms within scientific inquiry, ranging from the unification of laws and theories across different domains of inquiry to the integration of diverse phenomena under overarching explanatory principles. Whether through the synthesis of disparate theories into a single overarching framework or the identification of common principles that underlie seemingly unrelated phenomena, explanations as unification offer a powerful means of simplifying and organizing scientific knowledge, providing a deeper and more coherent understanding of the natural world.

Central to the concept of unification in science is the idea of explanatory coherence, which entails the integration of diverse empirical observations and theoretical principles into a unified and internally consistent framework. Explanations as unification seek to identify commonalities and connections that link seemingly disparate phenomena, thereby illuminating the underlying unity that pervades the diversity of the natural world. This emphasis on coherence reflects a commitment to simplicity and elegance in scientific explanation, privileging theories and frameworks that offer the most parsimonious and comprehensive account of empirical data. By unifying diverse phenomena under a

single explanatory framework, scientists can simplify and systematize their understanding of the natural world, providing a more cohesive and integrated picture of reality.

One of the key virtues of explanations as unification lies in their ability to reveal deeper patterns and regularities that transcend the specific details of individual phenomena. By identifying commonalities and connections across different domains of inquiry, unification allows scientists to uncover underlying principles that govern the behaviour of complex systems. This emphasis on underlying unity enables scientists to move beyond mere description and classification to uncover deeper causal relationships and explanatory structures that unify diverse phenomena. For example, the unification of electromagnetic and gravitational forces under the framework of general relativity revealed a deep connection between seemingly distinct physical phenomena, paving the way for a more unified understanding of the fundamental forces that govern the universe. Similarly, the unification of genetics and evolutionary biology under the framework of the modern synthesis provided a unified account of the mechanisms of inheritance and variation, offering insights into the processes of adaptation and speciation that underlie the diversity of life.

In addition to their explanatory power, explanations as unification also offer a number of methodological and epistemological benefits within scientific inquiry. By providing a

coherent and integrated framework for organizing empirical data and theoretical principles, unification facilitates the discovery of new phenomena and the formulation of novel hypotheses. Scientists can use the principles of unification to identify gaps and inconsistencies in existing theories, thereby guiding the development of new research programs and experimental investigations. Moreover, unification fosters interdisciplinary collaboration and communication by providing a common conceptual framework that transcends disciplinary boundaries. By promoting the exchange of ideas and methodologies across different fields of inquiry, unification fosters innovation and cross-fertilization, leading to new discoveries and insights that transcend the limitations of individual disciplines.

Despite its many virtues, explanations as unification also face a number of challenges and criticisms within the philosophy of science. One of the key challenges lies in the tension between simplicity and complexity in scientific explanation. While unification seeks to uncover underlying patterns and regularities that simplify and organize scientific knowledge, it must also contend with the inherent complexity and diversity of the natural world. Complex systems often exhibit emergent properties and nonlinear dynamics that defy simple reductionist explanations, posing challenges to the project of unification. Moreover, the pursuit of unification can sometimes lead to oversimplification or reductionism, glossing over important nuances and complexities

in favour of a more unified but less accurate account of reality. As a result, scientists must strike a balance between simplicity and complexity in their quest for unification, recognizing that while simplicity is a virtue in scientific explanation, it must always be tempered by a commitment to simplicity and informative ability.

Another challenge facing explanations as unification lies in the identification and evaluation of unifying principles themselves. While the pursuit of unification is guided by a desire to uncover underlying patterns and connections that link disparate phenomena, the identification of these principles is often fraught with uncertainty and ambiguity. Unifying principles must be sufficiently general and abstract to apply across diverse domains of inquiry, yet sufficiently specific and explanatory to provide genuine insight into the underlying structure of reality. Moreover, unifying principles must be empirically adequate, meaning that they must accurately capture the empirical data and observational evidence without ad hoc modifications or theoretical embellishments. As a result, the identification and evaluation of unifying principles require careful empirical investigation and theoretical analysis, often involving iterative cycles of hypothesis testing and revision.

In conclusion, explanations as unification represent a fundamental aspiration in the philosophy of science, embodying the quest to reveal underlying patterns and connections that unify seemingly disparate phenomena. By integrating diverse empirical

observations and theoretical frameworks under a unified explanatory framework, unification offers a powerful means of simplifying and organizing scientific knowledge, providing a deeper and more coherent understanding of the natural world. However, the pursuit of unification also faces challenges and criticisms, including the tension between simplicity and complexity in scientific explanation, and the identification and evaluation of unifying principles themselves. Despite these challenges, explanations as unification remain a central goal of scientific inquiry, guiding the search for deeper principles and regularities that govern the complexity of the natural world.

1.2.6. Pragmatic Model of Explanation

The pragmatic model of explanation was developed by Van Fraassen (1980). According to his view and his antirealism, a theory does not need to be true to offer an explanation of an event *E*. Moreover, he does not want to build a model of the scientific explanations only, but in general his aim is to analyse each kinds of why-questions and their answers. According to him, there are not two objects (explanations and facts), but three: explanations, facts and contexts. Moreover, his model is built upon Bromberger's (1966) elaboration: a why-question *Q*, which is an

interrogative about an object or a set of objects in a specific context, and it can be conceived as a triple: $Q = (P_k, X, R)$ where:

1. P_k is the question theme;
2. $X = (P_1 \ldots, P_k, \ldots)$ is the class of antithesis;
3. R is the relation of explicative relevance.

Let's take as example the why-question "Why Romeo killed himself?". As to analyse the question in Van Fraassen manner, we distinguish three issues: 1) the theme of the question, 2) the antithesis classes, and 3) the relevance relation *R*. The 1) concerns the action (suicide) Romeo committed. The antithesis classes are the set of the different interpretations of the question. They vary accordingly to the attention we pose to the different items of the question: P_1 – Why *Romeo* killed himself?; P_2 – Why Romeo *killed* himself?; P_3 – Why Romeo killed *himself*? The last, the relation of explicative relevance regards the point of view of the inquirer, looking for the definition of the reasons at the basis of the action, the phenomenon, the object to explain.

In conclusion, there is the question *Q* and the core of the answer *A*, a proposition that could result relevant in virtue of its relation with the couple $Q = (P_k, X)$. The relevance of *A* respect to *Q* is to be determined in virtue of its relevance respect to the couple according to *R*. It means that we could receive two different answers (or more) from the same question. If we ask

Juliet “Why did Romeo die?”, we receive a different answer than the one we could have from the coroner after the examination of his body. Both answers are correct, but their relevance differ accordingly to *R*, based on different contexts.

The general scheme of an answer *B* to *Q* is a proposition of that kind:

B = because *A*

In this schema, *A* is the nucleus of the answer, and *B* is related to the relevance relation varying accordingly with the context. According to Van Fraassen (1980) a why-question has three main assumptions: 1) the question theme is true; 2) except the theme, each element of its antithesis class is false; 3) it is at least true one of the propositions *A* that are related with the theme and the antithesis class according to the explanatory relevant relation.

With this approach Van Fraassen proposes to overcome the problems other theories of explanation encounter. The main feature of the pragmatic approach is the focus on the context. With this model, Van Fraassen opens a new way to account for scientific explanations, different from the syntactic and semantic framework grounding the previous theories. The pragmatic approach reveals that to give an explanation does not consist only in the application of an explanatory model to a specific case, but it takes into account the knowledge, the intentions, the assumptions

of both the inquirer and the inquired. To ask for an explanation is to require an answer to a question, and the answer depends on the context in which it is made and the knowledge of who answers. The importance of context will be a crucial feature in the De Regt's theory of scientific understanding, as we will see in Chapter 4.

1.3. Open Issues

Upon examining recent efforts to construct models of scientific explanation, it becomes evident that they yield both significant insights and lingering questions, laying the groundwork for future research avenues. While acknowledging the inherent subjectivity of any assessment, certain key observations emerge, although subject to ongoing debate.

A central issue revolves around the role of causation in scientific explanation. While traditional models aim to capture causal relationships, challenges such as explanatory asymmetries and irrelevance suggest existing models may fall short. For example, the Statistical Relevance (SR) model has been criticized for its limited ability to convey causal information solely through statistical relevance. Similarly, the Causal Mechanism (CM) model struggles to fully delineate causal relevance relations. This underscores the need for a more nuanced understanding of

causation, urging integration with the expanding literature on causation.

However, it is essential to consider whether an exclusive focus on causation overshadows other essential aspects of explanation. While causation is crucial, not all causal claims offer equally profound explanations. Variations in the generality and coherence of causal claims highlight the need to discern what constitutes a "good" causal relationship for explanation, suggesting that traditional concerns extend beyond mere causal distinctions.

Moreover, the debate extends to the existence and interaction of non-causal forms of explanation with causal explanations. Recent interest suggests that non-causal explanations may be prevalent in science, prompting inquiries into their structure and relationship with causal explanations. This raises the question of whether a unified theory can encompass both causal and non-causal explanations.

Another critical consideration pertains to the pursuit of a singular model of explanation. While universalist models aspire to offer a unified framework applicable across disciplines, disciplinary differences in explanatory practices pose challenges.

Furthermore, models imposing specific requirements on explanation may not universally apply across disciplines. Disciplines like biology and social sciences may not conform to models centred on discovering laws, and the feasibility of certain

models may vary due to inherent differences in data availability and interpretive frameworks.

In conclusion, recent discussions on models of scientific explanation underscore the pivotal role of causation while urging a broader consideration of explanatory factors. Future endeavors should aim for a nuanced understanding of causation, explore the coexistence of causal and non-causal explanations, and develop discipline-sensitive models to enhance our appreciation of scientific explanation across diverse domains. In particular, to the extent of this thesis, I want to stress the importance of further studies about the role of explanations and understanding in the field of Artificial Intelligence, and specifically how will be crucial to future research the study of Deep-learning models, as we will read in the next Chapters 7 and 8.

1.4. Understanding and Scientific Understanding

In this section, I present briefly the main themes of the epistemology of understanding and the relevant aspects of scientific understanding. Understanding has been conceived in the XIX Century as the specific outcome of the humanities and specifically the social sciences. Its conception is a distinction from the output of the natural sciences: explanations. From the notion of explanation we can derive the notion of understanding as

Verstehen. Now the distinction is not so sharp and clear[14] and a specific epistemology of understanding has been proposed by several scholars.

Scientific enterprise, as a whole of many practices, methodologies, and a family, wide and broad, of different scientific communities, is also made of reports, papers, and books in which it has often used the word "understanding". For example, in his wonderful and very communicative book published in 2015, Matthew Cobb, British zoologist and Professor of zoology at the University of Manchester, in *Life's greatest secret*, after the presentation of the article of Jim Watson and Francis Crick (1953) in which they advance the new category of genetical information, writes:

This was a category that Watson and Crick had just invented. And yet it was familiar because it fitted so well with the ideas that were in the air at the time. It was adopted without debate; this new way of looking at life seemed so obvious that it was immediately accepted by scientists around the world. Today, these words, or something like them, are said every day in classrooms all over the planet as teachers explain the nature of genes and what they contain. This book explores the surprising origin of these ideas, which can be traced back to physics and mathematics, and to wartime work on anti-aircraft guns and signals communication. It describes the way in which these concepts entered biology through the then-fashionable field of

[14] See Hacking (2001), *Lectures at the Collège de France* (2000-2006), specifically the lectures *Façonner les gens I* and *II* (Hacking, 2001).

cybernetics, and how they were transformed as biologists *sought to understand*[15] life's greatest secret – the nature of the genetic code (Cobb, 2015: 3).

Understanding needs new categories to describe the new findings of researchers. After a couple of year, in December 2019, Cobb writes a book about the brain and the history of the science of it, the first organ we actually use to do science. At the very beginning of the book, the traces back the history of this scientific research to Nicolaus Steno, anatomist from Denmark:

In 1665 the Danish anatomist Nicolaus Steno addressed a small group of thinkers gathered together at Issy, on the southern outskirts of Paris. This informal meeting was one of the origins of the French *Académie des Sciences*; it was also the moment that the *modern approach to understanding* the brain was set out. In this lecture, Steno boldly argued that if we want to *understand what the brain does and how it does it*, rather than simply describing its component parts, we should view it as a machine and take it apart to see how it works. […] Although most neuroscientists have never heard of Steno, his vision has dominated centuries of brain science and lies at the root of *our remarkable progress in understanding* this most extraordinary organ[16] (Cobb 2020: 14).

Understanding is manifold; one kind can be described as objectual understanding, understanding what; it is when we say "look, I

[15] Italics are mine.
[16] Italics are mine.

have understood *what* the brain does and *how*". It can also be conceived as explanatory understanding, the understanding why something is the case, i.e. why the brain (even if not the whole brain) is needed to have binocular sight. Likewise, let's take the example of the James Webb telescope, which images released since July 2022 made us hold ours' breath. Webb telescope has the high technological ability to *see back* into time thanks to infrared mirrors: «Webb is able to see back to about 100 million - 250 million years after the Big Bang. But why do we need to see infrared light to *understand* the early universe? Because light from these objects is shifted to the red»[17].

These are the two main forms of understanding which epistemologists study, the ones arguing for the irreducibility from one form to another, the others defending the reducibility of them, some others arguing for the distinctive nature of understanding as not know-how[18]. The form of understanding Cobb is thinking about is the one strictly connected to representations and it is the kind of scientific understanding one can achieve using models, as preferential example of scientific representation. I will argue in the following chapters that there is a distinctive objectual understanding that the main two views we will see, De Regt's and Khalifa's theories, cannot account to. I think these views should be ameliorated

[17] See https://webb.nasa.gov/content/science/firstLight.html. Credit, Aleš Tošovský.
[18] See for example, Sullivan (2017). For more broad discussion about this issue, see Grim, Baumberger and Ammon (2017) and Grimm (2021).

in order to find space also for this kind of understanding, which in certain cases, such as deep-learning models, as I will describe in Chapters 4 and 5, cannot be reduce to understanding why. On the contrary, some understanding scholars consider only objectual understanding as specifically valuable. It is said to concern wide subject matter that involve «a large network of propositions and relations between those propositions» (Carter and Gordon, 2014: 8). But against it, I will support De Regt's and Khalifa's effort to develop and epistemology of explanatory scientific understanding. My suggestion, in addition to their work, is that it is worth taking also the objectual understanding, as representational understanding under the aegis of the understanding scholarship.

The objectual understanding (OU) is conceived in the epistemological literature as the principle form of understanding, but there is no consensus on that. Carter and Gordon are ones of the strongest scholars arguing for the centrality of objectual understanding: «We think it is clear that objectual understanding - for example, as one attains when one grasps the relevant coherence-making relations between propositions comprising some subject matter - is a particularly valuable epistemic good» (Carter and Gordon, 2014: 7-8); and we have already seen as OU is analysed also in terms of the value it carries in the epistemological vocabulary.

Moreover, understanding comes progressively. While knowledge is a static outcome, understanding is a process that nevertheless conduces to a more or less structured account (which involves

knowledge, i.e. scientific knowledge) of a phenomenon. We can find degree of understanding in familiar cases such the learning of a foreign language; during the first years of school we will have a minimal understanding of it, while through study and exercise the understanding will increase and the wider it gest the more it become cognitively demanding. In fact, understanding «wider subject matters will tend to be more cognitively demanding than understanding narrow subject matters because more propositions must be believed and their relations grasped» (Carter and Gordon, 2014: 7-8). The same we can observe obtains in cases of scientific understanding, in which historically, as the Cobb's example, we notice an increase of understanding of a phenomenon, thanks to the progress of science. Indeed, Cobb focuses also on the history of science, which could be refined as history of understanding:

> Understanding how past thinkers have struggled to understand brain function is part of framing what we need to be doing now, in order to reach that goal. Our current ignorance should not be viewed as a sign of defeat but as a challenge, a way of focusing attention and resources on what needs to be discovered and on how to develop a programme of research for finding the answers (Cobb 2020: 14).

Shifting from the diachronic discussion about scientific understanding, we find the synchronic issue about the relation between understanding and knowledge, which relates to the relation between understanding and truth. This is where the scholars disagree

about the factivity of understanding. The issue concerning the factivity of knowledge has been well debated in epistemology. Factivism about knowledge requires the knowers to have beliefs in true propositions. Factivists about understanding, therefore, argue that understanders need to have beliefs in true explanatory propositions, i.e. "*q* explains why *p*". This holds, if we accept that for an agent to understand *p* implies to know that *q* explains *p*. If we take the side of the factivists, then, given the factivity of knowledge, also to know that *q* explains *p*, requires the agent to posses belief in the proposition that "*q* explains *p*" and this belief is true.

On the opposite, the non-factivists doubt that understanding tracks the truth (De Regt 2015, 2017; Elgin, 2004, 2007, 2009a, 2009b; Riggs, 2008; Zagzebski, 2001). They can be the sort of scholars focusing on explanatory understanding, but claiming that understanding is decoupled from knowledge and therefore it does not require for an agent to possess true belief of explanatory propositions.

There is a *tertium* stance here, an alternative to the polarized views of factivism and non-factivism, which is the quasi-factivism. Scholars of this view deny the tenets of non-factivists and hold that understanding and truth are firmly connected (Greco, 2013; Grimm, 2006; Knanvig, 2003; 2009a; Mizrahi, 2012; Pritchard, 2007). There is an interpretation of quasi-factivism given by Khalifa (2023), *voluntarism*, which leaves

open the door for both realist and anti-realist stance towards understanding.

The debate about scientific understanding inherits these themes and related problems, but with the specific focus on scientific activities, which are mainly based on questions, going up and down to discover, explain, describe, falsify or testify ideas, hypotheses, theories and models. Scientists need conceptual frameworks, advanced methodologies, technologically equipped tools, and much discussion in order to achieve an understanding of a phenomenon.

Chapter 2:
Verstehen and Understanding

2.1. Introduction to *Verstehen*

One of the first testimonies of the distinction between understanding in the humanities and the natural sciences comes from History. Namely, Droysen suggested that history requires understanding while physical sciences explain (Droysen, 1868). This definition has held for many decades, but recently, thanks to the epistemological research about understanding, this distinction has been criticised. This is the case of Ian Hacking's (2002, 2004) lectures at College de France entitled *Façonner les gentes*, in which Hacking talks about the definition of the core idea at the basis of the distinction between sciences and humanities, which collapses in the cases he studies, i.e. disability. The use of categories from the sciences ends up with a categorial consequence on the persons they apply to. This consequence crushes the main distinction between the concepts of sciences and the concepts of the humanities. This is what

has been called *Verstehen* (in German). The same intricacies of defending the claim about distinctiveness between understanding in the natural sciences and *Verstehen*, conceived as understanding in the human sciences, are discussed by Khalifa (2019). His main argument could be taken as an implementation of Hacking's argumentation. He notices that many scholars have argued that *Verstehen* is a unique form of understanding that the natural sciences do not have. He denies this sharp distinction; better, he denies there is a distinction and argues that *Verstehen* in human sciences is a species of scientific understanding (Khalifa, 2019).

But the origins of *Verstehen* have to be settled within the discussion of German historicism and the tendency of the sociologists to distinguish the kind of understanding they can achieve of sociological phenomena they study, from the one of the natural sciences. Distinction that Hacking aims to criticize.

To sum up the challenges of understanding, I find this passage by Danielsson a beautiful summary of what makes us as humans aspire to understand:

> All the things we can see and experience, our material universe as well as life and consciousness, are different aspects of the same world. The task of science is to find out as much as possible about how this, our universe, works. It does so with the help of various disciplines such as physics, chemistry, and biology, which through medicine and neuroscience translate into the humanities, social sciences, and philosophy. They all have their own language and their own criteria for truth. We see these different parts

of knowledge as more or less independent of one another. […] Our organic bodies, all our thoughts, including the scientific models we create, are parts of the same world that we so desperately want to *grasp* [emphasis mine]. The physics I imagine must handle everything; nothing must be left aside. (Danielsson, 2023: 15).

The physicist Danielsson imagines is an intertwined collaborative work of scientists-understanders that cooperate with each other. As understanding is contextual to a mind, a brain, a body and the social community in which scientists work, Danielsson stresses the focus on the urgency to understand how human understanding works, in order to figure out how it is possible and what is to understand the world, the cosmos and their phenomena scientifically. If the task of science, as Danielsson claims (2023), is to find out (as much as possible) how our universe works, we have to define in the first place what "our universe" refers to. According to him, it is to be done with the help of many disciplines, natural and human sciences, pure and applied sciences and humanities. From this line of thought takes form the idea that the connection between the many sciences is helpful to understand *our* universe, from the scientific point of view, which is a plural one. This connection of disciplines reminds to the later conflicts between human sciences and natural sciences in the XIX Century, in which the notion of understanding was relegated to the domain of the human-socio-historical sciences (or also social science), while the notion of explanation was in play in the natural sciences. The human-socio-historical science, was claimed, aims at

understanding (*verstehen*, verb) the human-socio-historical phenomena, while natural sciences aim at explaining phenomena.

Verstehen is a term used in the philosophy of social science which involves «the grasping of meaningful behaviour, psychological states, practices, and cultural artifacts» (Khalifa, 2019: 282). Traditionally, there are two sides of the discussion: the hermeneuticists, now "interpretivists" (Stueber, 2006), taking *Verstehen* as to distinguish the social from the natural sciences, and the positivists, now "naturalists" (Roth, 2003), who declined such distinction. Now, there are still scholars defending the difference between the two kinds of sciences, while Khalifa (2019), on the other hand, argues that all the norms that govern understanding in the natural sciences also apply to understanding in the human sciences. So, I will take Khalifa as one of the proponent of the continuity view between *Verstehen* and scientific understanding, while I will describe Dilthey's view as the traditional hermeneuticist arguing for the distinctiveness of *Verstehen*.

2.2. Dilthey on *Verstehen*

In the second half of the XIX century, Dilthey introduced the notion of *Verstehen* as a foundational principle in the social sciences, particularly in his hermeneutic approach to understanding human behaviour and culture. *Verstehen*, which translates to

"understanding" or "comprehension" in English, is central to Dilthey's methodology for interpreting and explaining human actions, experiences, and expressions within their historical and cultural contexts. Around 1883, Dilthey published the first and last volume of *Einleitung in die Geisteswissenshaften* (Introduction to the human sciences), in which he wanted to defend the autonomy of socio-historical sciences (human sciences or *Geisteswissenshaften*), from the natural sciences. He aspires to define their specific features and the conditions under which the human sciences can gain soundness. Using the categories of "hermeneuticists" and "positivists", we can define Dilthey as a hermeneuticist, as he distinguishes the human and natural sciences. According to him, in the first place, the two sciences differ from the object of study. The object of *Verstehen* is the domain of the human beings, in which we are agents and of which we have awareness. On the other hand, the object of natural sciences is the amount of phenomena independent from the human minds. In the second place, there is also an epistemological difference between the sciences: according to Dilthey the data of the natural sciences are the result of the observation of the outer world, while the data of the human sciences come from the *Erlebnis*, the inner experience, and the understanding an individual can have of another human being. The third difference can be seen in the kinds of categories the two scientific areas use: the natural sciences produce explanations (of different sort); the human sciences use distinctive categories as meaning, purpose and value.

These differences concern the relation between the subject and the object of inquiry. In the natural sciences the subject and the object are heterogeneous. In the human sciences, the subject of inquiry is also its object. According to Dilthey the world of the human-beings has its core in the individual and it is structured by relations historically bound, from which the cultures and the social organizations arise. The structure of the human world is historical, so – Dilthey argues – individuals and human structures have to be understood in their historical features. The Dilthey's *Einleitung* (1883) signs the starting point of the German historicism. According to him, the historical research focuses on the manifestation of the human world in the individualities, while the human sciences, such psychology or anthropology, aim to figure out the invariances of the human world.

According to Khalifa (2019: 284), *Verstehen* is characterized as «some kind of accurate simulation of another's mental states and processes, and human-scientific accounts – most notably intentional-action explanations – improve in proportion to the *Verstehen* they recruit, enable or provide». This idea was embedded already in Dilthey's account, according to which *Verstehen* represents more than just a cognitive act of grasping facts or events; it involves a deeper empathetic engagement with the subjective experiences, intentions, and meanings underlying human actions. He emphasized the importance of empathy, imagination, and intuition in the process of understanding others, recognizing that individuals' actions are

shaped by their unique perspectives, beliefs, and emotions. Thus, Verstehen requires researchers to adopt a holistic approach that considers the subjective dimensions of human life, rather than reducing phenomena to mere observable facts or statistical data.

Dilthey's conception of Verstehen was a response to the positivist tradition, which sought to apply the methods of the natural sciences to the study of human behaviour. He argued that human beings are fundamentally different from natural objects, and therefore, the methods of natural science are inadequate for fully understanding the complexities of human experience and culture. Instead, Dilthey proposed a hermeneutic approach that emphasizes interpretation, dialogue, and the reconstruction of meaning.

In Dilthey's view, *Verstehen* is not only a methodological tool but also a philosophical stance that acknowledges the intrinsic value of human subjectivity and the diversity of cultural expressions. By engaging in Verstehen, researchers can gain insight into the lived experiences of individuals and communities, uncovering the underlying motives, values, and symbolic frameworks that shape human behaviour and social institutions.

Overall, Dilthey's conception of *Verstehen* represents a significant departure from the positivist approach to social science, offering a richer and more nuanced understanding of human life through the cultivation of empathy, interpretation, and dialogue.

2.3. Khalifa on *Verstehen*

In Chapter 5, I will address Khalifa's account of scientific understanding, but here we can insert some of Khalifa's ideas to foster the discussion about *Verstehen*. Khalifa (2019) argues that the concept of understanding in the human sciences (*Verstehen*) is not fundamentally different from the concept of understanding in the natural sciences, contrary to the traditional dictum. In fact, as we have seen, conventionally, philosophers have debated whether *Verstehen* allows the human sciences to achieve a unique kind of understanding not found in the natural sciences. Khalifa (2019) proposes to call those who claim that *Verstehen* is distinct from scientific understanding "interpretivists[19]", and those who defend the indistinctiveness between *Verstehen* and scientific understanding "naturalists".

Khalifa (2019: 283) proposes a new way to frame this debate:

> Traditionally, when philosophers of social science have debated about Verstehen, only interpretivists lay claim to a concept of understanding. However, current work in epistemology and the philosophy of natural science challenges this framework by offering more general concepts of understanding, intended to encompass both the human and natural sciences. These developments shift the terms of the debate. No longer is the question whether the human sciences alone aim at understanding.

[19] I suppose the coin recalls the hermeneutic *niveau* in which the German historiographical theorization flourished especially thanks to Dilthey's work.

Rather, the question is whether and to what extent the human and natural sciences aim at the same kind of understanding.

He argues that both the human sciences and the natural sciences aim for understanding, but the question is whether *Verstehen* fulfils the same criteria for understanding as scientific understanding.

Khalifa outlines his Explanation-Knowledge-Science (EKS)[20] model of understanding, which has three core principles:

a) *Minimal Understanding*: Belief in an approximately true explanation of a phenomenon.
b) *The Nexus Principle*: Understanding improves as we grasp more explanatory information about the phenomenon.
c) *The Scientific Knowledge Principle*: Understanding improves as our grasp of the explanation resembles scientific knowledge, achieved through a process of considering and comparing potential explanations using the best available methods and evidence.

Khalifa applies the EKS model to two examples from social science research: a study on why students avoid seeking help in school; and an ethnography of headhunting practices among the Ilongot tribe. In particular, the case study of Ilongot headhunting practices serves as an example of how the EKS model can capture the process of

[20] I will describe in details the EKS model in Chapter 5.

understanding a phenomenon through the anthropological lens of Renato Rosaldo (1980). In both cases, Khalifa argues that the researchers demonstrate understanding according to the EKS Model. This suggests that *Verstehen* can be a form of scientific understanding.

In conclusion, Khalifa acknowledges that more research is needed, but concludes that interpretivists (those who argue for a unique kind of understanding in the human sciences) face a difficult challenge. They would need to show that *Verstehen* violates one of the principles of the EKS model, while still being a valid form of understanding in the natural sciences.

2.4. Differences and Connections in Dilthey and Khalifa's *Verstehen*

Kareem Khalifa's and Wilhelm Dilthey's conceptions of *Verstehen* offer two distinct perspectives on the nature of understanding within the realm of social sciences and humanities. Khalifa's articulation, particularly in his paper on the notion of *Verstehen* as scientific understanding, challenges the traditional view espoused by Dilthey, who posited that verstehen, or understanding, was fundamentally different from explanation. Khalifa's approach seeks to bridge the perceived gap between verstehen and explanation, arguing for a more integrated and scientific

understanding that incorporates elements of both. In contrast, Dilthey emphasized the unique character of verstehen, suggesting that it involves a form of empathetic understanding that cannot be reduced to causal explanation. Understanding these differing perspectives requires a deep dive into the philosophical underpinnings of both Khalifa's and Dilthey's ideas.

Wilhelm Dilthey, a prominent German philosopher of the late 19th and early 20th centuries, is often credited with articulating the concept of *Verstehen* as central to the human sciences. For Dilthey, *Verstehen* entails a process of empathetic understanding that allows one to grasp the subjective meanings inherent in human actions and expressions. Unlike the natural sciences, which seek to explain phenomena through causal laws, Dilthey argued that the human sciences require a hermeneutic approach that acknowledges the unique context and lived experiences of individuals. According to Dilthey, this form of understanding involves entering into the mindset of the subject, discerning the intentions and motivations behind their actions, and interpreting the significance of their expressions within the cultural and historical framework in which they occur. In essence, *Verstehen*, for Dilthey, is a deeply interpretive endeavour that recognizes the complexity and richness of human existence.

However, Kareem Khalifa's perspective on verstehen challenges Dilthey's notion by proposing a more scientifically oriented understanding that seeks to incorporate elements of explanation

within the framework of interpretation. Khalifa argues that while Dilthey's emphasis on empathy and interpretation is valuable, it does not fully capture the potential for a more rigorous and systematic approach to understanding human phenomena. In his paper, Khalifa suggests that a scientific understanding of verstehen can be achieved by integrating methods and insights from both the natural and human sciences. Rather than viewing verstehen as fundamentally distinct from explanation, Khalifa proposes that a more fruitful approach involves recognizing the complementary relationship between the two.

According to Khalifa, scientific understanding in the context of verstehen entails not only interpreting the meanings and intentions behind human actions but also seeking to uncover the underlying causal mechanisms that govern these actions. This approach involves formulating hypotheses, testing them through empirical observation and experimentation, and refining our understanding based on the evidence gathered. By adopting a scientific methodology, Khalifa argues that we can achieve a more objective and systematic understanding of human behaviour, one that is grounded in empirical evidence rather than subjective interpretation alone.

Central to Khalifa's argument is the idea that scientific understanding does not entail reducing human phenomena to deterministic causal laws but rather recognizing the complex interplay of multiple factors that contribute to human behaviour. In this sense, Khalifa's conception of verstehen as scientific

understanding retains the interpretive dimension emphasized by Dilthey while also incorporating elements of causal explanation derived from the natural sciences. By embracing a more interdisciplinary approach that draws on insights from psychology, sociology, anthropology, and other fields, Khalifa suggests that we can develop a more comprehensive and nuanced understanding of human behaviour.

However, despite the differences between Khalifa's and Dilthey's conceptions of verstehen, there are also points of convergence between the two perspectives. Both acknowledge the importance of context in understanding human actions, recognizing that meaning is inherently tied to the cultural, historical, and social conditions in which it emerges. Additionally, both Khalifa and Dilthey emphasize the need for reflexivity and self-awareness in the process of understanding, acknowledging the role of the observer's own perspectives and biases in shaping their interpretations.

In conclusion, Kareem Khalifa's conception of verstehen as scientific understanding offers a compelling alternative to Wilhelm Dilthey's traditional view, challenging the perceived dichotomy between interpretation and explanation. By advocating for a more integrated and interdisciplinary approach that incorporates insights from both the natural and human sciences, Khalifa seeks to develop a more rigorous and systematic understanding of human behaviour. While Khalifa's perspective may diverge from Dilthey's in certain respects, both share a common commitment to exploring the

complexities of human existence and uncovering the underlying patterns and mechanisms that govern it. Ultimately, understanding the nuances of verstehen requires engaging with the diverse philosophical perspectives that have shaped its conception over time, from Dilthey's hermeneutic tradition to Khalifa's scientific orientation.

Chapter 3:
Schurz and Lambert on Scientific Understanding and the Received View

3.1. Schurz and Lambert's Theory of Scientific Understanding

The philosophical debate between explanation and understanding in the XX Century set explanations as winners, as we have seen. It was only in the 1990s that the notion of scientific understanding had a turnover. Schurz and Lambert are the first ones in 1994 to write extensively about scientific understanding. The primary task of their essay is to outline a [basic] theory of scientific understanding:

> It is in terms of this theory that the relation between scientific understanding and scientific explanation shall be ascertained. Two constraints help to define this enterprise. First, the definition of scientific understanding, as a notion to be developed, presumes that scientific understanding is an intersubjective (objective) notion and independent of

the psychological features of given thinking humans. Second, to avoid any hint of circularity in the relation between scientific explanation and scientific understanding, the explication of scientific understanding will be *independent* (Schurz and Lambert, 1994: 66).

Formalization of unification is their candidate as general theory for the many kinds of explications. According to them, understanding a phenomenon *P* is a kind of fitting in a cognitive corpus (UFC):

(*UFC*): To understand a phenomenon *P* – for a determined agent *S* – is to insert *P* in the cognitive corpus *C* of *S*.

In this formulation, *C* is the cognitive corpus[21] of the inquirer, it contains all statements known or believed by the inquirer (including observation statements, laws, theories and hypotheses of varying degree of belief): «a phenomenon is contained in *C*, or fits into *C*» literally means «the statements expressing this phenomenon is contained in *C*, or fits into *C*» (Schurz and Lambert, 1994: 68). There is a threefold relation: 1) answer to question demanding understanding, 2) Phenomenon, and 3) C, cognitive corpus of an inquirer. They interpret this relation within the logic of answers and questions:

[21] Schurz and Lambert do not distinguish ontologically and referentially between the scientist agent *S* and their cognitive corpus *C*: «Just as it makes sense to say in abstracto that the brain of an agent (rather than merely the agent) understands, so it makes sense to say in abstracto that C understands» (Schurz and Lambert, 1994: 76).

Def. 1: statement *A* contributes understanding of *P* to *C* iff *A* is an adequate answer to the question "How does *P* fit into *C* + *A*?", where the phenomenon *P* is contained in the cognitive corpus *C*.

The conditions to be satisfied by this definition are the following:

1) contributing understanding of *P* presupposes that *P* is always known or believed in *C*;
2) the notion of contributing understanding assumes that the answer *A* is not yet known in *C*;
3) *C* = <*K*, *I*> stands for the cognitive corpus of some idealized agent at some time. The corresponding agent may be a person, a scientific community an intelligent machine such as computer.

In this formulation, *K* is the set of (elementary) phenomena known or believed by the inquirer at a given time; *I* is a set containing the arguments mastered by the inquirer at a given time. So, Schurz and Lambert picture the understanding as a kind of inferential relation of question-answer in a cognitive corpus, which can add new information. In order to be *P* understandable, an answer *A* needs to add new information to *C*: may the information be descriptive, they could be added to *C* (in that case *A* is factive or theoretically innovative, according to their terminology). Moreover, *A* can otherwise contribute as new inferences that can be added to *C*, given

that it has not been known previously or mastered by *C*, proving that *P* is inferable from *C*.

Nonetheless, according to Schurz and Lambert it is not sufficient to define the connections between *P* and *K* elements, in order to insert correctly *P* in $C = \langle K, I \rangle$. This means that *C* must be coherent. To avoid uncoherent error, they use the Coherence Principle (Coer):

Coer: to fit *P* into *K** means (informally) to connect *P* with parts of *K** such that the coherence of *K** is increased relative to the coherence of *K*. The total coherence of *K** increases just in case the local coherence resulting from the connection of *P* with parts of *K** is not outweighed by a loss of coherence elsewhere in *K** (Schurz and Lambert, 1994: 71).

Systems of scientific knowledge are for them informative systems represented linguistically. Connections between the elements of *K* are arguments. The correctness of an argument cannot be defined by a sufficient condition – they offer examples of arguments which define extensively the correctness of a scientific argument. The critical consequence of this claim by them is that many theories of explanation, may be nomologic, or causal theory, are characterized locally and therefore they are not suited to represent the global feature of explanation as in *Coer*. On the other side, each argument contributing understanding which is not linked to any causal argument would not be considered an explanation. In conclusion,

they are committed to a pluralist conception of explanation, according to which kind of element of *C* are involved and which kind of information are necessary to add *P* in *C*.

Even if the accounts of scientific understanding advanced later on by De Regt and Khalifa differ from the one proposed by Schurz and Lambert, they reveal the relevant purpose of the scholars of understanding, which is to find a way in which unification is made possible to support an account of scientific understanding. Moreover, they employ all the elements on which De Regt and Khalifa will work: the (skill/ability) cognition of the agent(s) and explanatory information.

De Regt admits transparently that his aim is to investigate and explicate the nature of the understanding that science can provide. First question: are there universal, timeless criteria for scientific understanding? His answer: even a cursory look at the history of science suggests – no –. He consider that the discussion about understanding in science had a central role in the genesis of quantum mechanics in 1920s – heated debates about the intelligibility of quantum theory and the question whether it can provide understanding of the phenomena in the domain of atomic physics. His claim: scientists' standards of intelligibility and understanding vary strongly – not only diachronically but also synchronically. That is why he focuses on skill, context and intelligibility of theories (De Regt and Dieks, 2005; De Regt, 2017), as we will see in the next chapter.

On the other side of the understanding story, there will be Khalifa (2017) as supporter of a new kind of explanatory understanding. His aim is to defend the view that understanding consists of knowledge of relevant explanatory information (an upgraded version of the received view).

3.2. The Received View about Scientific Understanding

Extracted from the work of Salmon (1989) and Lipton (2004), for many years the received view of understanding basically claims that understanding (but we can also apply it to scientific understanding) is to know causes, which can be conceived as explanatory knowledge. The received view (RV) holds that understanding is explanatory knowledge:

RV: *S* understands why *p* if and only if there exists some *q* such that *S* knows that *q* explains why *p* (Khalifa, 2023: 18).

From this definition understanding requires necessarily explanatory knowledge possessed by an agent *S* of a phenomenon *p*; and it says that to understand *p* is sufficient to have explanatory knowledge of *p*. The discussion about the theories of explanation from the XX Century is still embedded in this definition and some issues are common. Of course the main opponents of the RV hold

that we can know an explanation by memory or passive testimony, without achieving understanding and tracking explanatory knowledge of *p* (De Regt, 2009; Hills, 2009; Knanvig, 2003; Newman, 2012, 2014; Pritchard, 2008, 2009, 2010). Khalifa (2017) on the contrary defends an implemented version of RV, which entails some of the traditional issues concerning ability, objectual issue, explanation, truth and luck. These are the following:

a) Ability issue: Do understanding and knowledge demand the same abilities or skills?
b) Objectual issue: Is objectual understanding the same as explanatory understanding?
c) Explanation issue: Does understanding require explanation?
d) Truth issue: Does understanding entail true belief?
e) Luck issue: Is understanding incompatible with epistemic luck?
f) General Value issue: Do understanding and knowledge have the same epistemic value?

On a first sight, proponents of the RV will be oriented in answering affirmatively to all these questions, while opponents to the RV, i.e. De Regt, will be much more critics. Each scholars working on understanding and scientific understanding should

answer these questions, in order to develop and set their account of understanding/scientific understanding. In the literature we find a dyadic dialectic which tenses between the pole of the realism or of the antirealism. This dyad is now under discussion, insofar as Khalifa and De Regt are interested in an exchange confrontation which may mediate towards a third way to address these issues. In the next chapters 4 and 5, I will present De Regt's and Khalifa's theory of scientific understanding.

Chapter 4:
The Contextual Theory of Scientific Understanding

The chapter delves into de Regt's contextual theory of scientific understanding, offering a comprehensive analysis of its key tenets and implications. De Regt's framework posits that understanding in science is context-dependent, challenging traditional notions of universal understanding. At its core, the contextual theory asserts that scientific understanding is contingent upon the explanatory context within which it operates. Through a nuanced examination of scientific explanations, de Regt contends that the intelligibility and relevance of theories play a crucial role in shaping understanding. The chapter explores how de Regt's contextual theory navigates complex philosophical terrain, addressing fundamental questions about the nature of scientific knowledge and explanation. By elucidating the contextual factors that influence understanding, the chapter sheds light on the dynamic interplay between theory,

explanation, and scientific practice. Moreover, it critically evaluates the strengths and limitations of de Regt's approach, highlighting its implications for contemporary debates in philosophy of science. Through a rigorous examination of de Regt's contextual theory, the chapter offers valuable insights into the nature and scope of scientific understanding, enriching our understanding of the epistemological foundations of science.

4.1. Introduction

After the attempt of Schurz and Lambert (1994) to define scientific understanding as a detailed concept in the philosophy of science, to be analysed with its own philosophical dignity, Henk de Regt and Dieks (2005) are the first to deal with the specific form of understanding sciences-oriented programmatically. In their first paper (2005), they argue that scientific understanding (SU) can be defined as a contextual theory, and its core is the general criterion of intelligibility of scientific theories.

One of the first scholars who decided to work on the notion of scientific understanding is Henk De Regt, who, together with Dennies Dieks, wrote in 2005 a paper about SU and its contexts, *A Contextual Approach to Scientific Understanding* (De Regt and Dieks, 2005). Thanks to their work, the notion of SU has attracted more attention through the years. We now can distinguish mainly two

areas of inquiry, which Khalifa (2023) defines friends of understanding (De Regt *in primis*), and frenemies of understanding (Khalifa as the main opponent of De Regt position). This first position about SU focuses on the contexts in which science is done, taking particular care to historical cases and procedures.

The conception of SU finds its place in a framework in which both De Regt and Dieks look at understanding as a peculiar good, in the sense that also other philosophers of understanding conceive its value_ «understanding seems a central good that we try to realize when we think about the world. More specifically, the value of understanding seems to surpass that of knowledge» recognize the understanding of nature as «one of the aims of science» (De Regt and Dieks, 2005: 137). One of their motivations is to work out the proverbial lack of consensus about what a scientific explanation has to be (Newton-Smith, 2000, 130-131). So they develop a unifying proposal, following the suggestion of Newton-Smith that the unifying concept of understanding should play the role of peace-keeper in the debate about explanation: «all explanations supposedly give understanding, and a general theory of understanding might tell us how» (De Regt and Dieks, 2005: 137). Nevertheless, in front of the proposal of the unifying concept of understanding, Newton-Smith takes the side of the traditional view that understanding has to be rejected, in so far as does not pass the scientific pedigree requirement, to be considered valuable for the science vocabulary, for its being a subjective, psychological and therefore unfruitful

concept (Hempel, 1965: 413). Against it, according to De Regt and Dieks SU can play the unifying role. A complete account of SU is beautifully exposed in a «superb book», as Suàrez defines it, which won the Lakatos Award in 2019, being a «magnificent example of how history and philosophy of science can be productively integrated», as the Lakatos Award Committee stated in 2019, it is *Understanding Scientific Understanding* (De Regt, 2017).

4.2. What is Scientific Understanding?

The first task to deal with is for De Regt and Dieks to picture broadly the boundaries of SU and tell mainly what is to understand a phenomenon scientifically. I think that we can describe the attempt to define SU as a philosophical activity routed at the central core of the philosophy of science. We have different examples of books aiming at elucidating the main tenets of philosophy of science, there is namely Vincenzo Fano's well-known manuscript entitled *Comprendere la scienza* (Fano, 2005)[22], as one of the aim of the general philosophy of science is to understand science, with the tools of philosophy. The perspective of De Regt, Dieks and the philosophers of understanding (and namely the philosopher of scientific understanding, that can be a subset of the formers), is one

[22] Fano's book (2005), in English, Understanding Science, is a beautiful introduction to epistemology of natural sciences, focusing on the main concepts of philosophy of science of explanation, scientific theory and scientific laws.

of a step back. The aim of their research is not only to understand science, which is already a hard task, but it is also to understand scientific understanding, namely to understand what is – how it works, what it takes – to understand phenomena scientifically.

Moreover, SU can be described as a subset of understanding (Grimm, 2021), which in general has not been a prominent topic in contemporary epistemology, being sometimes neglected or suspected carrying a cheap philosophical bargain. Its recovery is recent and dates back between the end of 90s and the beginning of 2000 (Elgin 1991, 1996; Zagzebski 1996, 2001). In the XIX Century understanding (in its German translation, *Verstehen*) played a crucial role in the distinction between humanities, *Geistwissenschaften*, and the sciences. The German historian Droysen distinguishes explanation from understanding, claiming that the history as a humanistic science calls for understanding (*Verstehen*), while physical sciences explain (*erlkären*), and here lies the traditional difference between the aim of sciences and the aim of humanities (Droysen, 1868). While Dilthey recognises the understanding as the specific achievement of the humanities (Dilthey, 1910: 98-100), the logical positivist began to contrast this sharp distinction, noticing a relation even feeble, between explanation and understanding (Hempel, 1942; von Wright, 1971; Stueber, 2012). Nonetheless, during the XX Century the primary notion in philosophy of science was explanation, to be analysed in terms of valid argument (Hempel and Oppenheim, 1948), while understanding was considered a

psychological byproduct or a pragmatic aspect of explaining (Hempel, 1965: 413)[23]. Only recently, SU has been established as a specific area of inquiry and characterized as a central goal of the sciences. In fact, De Regt and Dieks (2005, 142) claim that «understanding is an inextricable element of the aims of science». In doing so, they follow the same idea of science some scientists have: «Science is the process that takes us from confusion to understanding in a manner that's precise, predictive and reliable» (Greene, 2008).

According to some philosophers of science (Friedman, 1974; Kitcher, 1981, 1989; Schurz and Lambert, 1994) the understanding achieved by science comes with a unified picture of the world. Some others argue that understanding comes with a causal connection, in particular as an output of causal explanations (Salmon, 1984, 1990, 1998; Humphreys, 1989; Dowe, 2000). Both the lists of authors, although, do not provide a complete characterization of understanding, simply affirming that a particular kind of explanation provides understanding, but no more.

De Regt and Dieks (2005) notice that the history of science gives us a wide number of examples in which what is called understandable varies across the time. Moreover, this variety shows how scientists across the history[24] disagree about what is

[23] See also De Regt (2009) for a taxonomy of relationship between explanation and understanding.

[24] See De Regt and Dieks (2005: 138): they use an example of differentiation in scientific criteria in order to have unanimous consensus about how to understand a phenomenon across time. Even if it is non-contentious that diachronically the consensus upon what criteria to be employed to catch understanding,

understandable and what is not. For example, we can find on one side Lord Kelvin's famous dictum: «it seems to me that the test of "Do we or not understand a particular subject in physics?" is, "Can we make a mechanical model of it?"» (Kargon and Achinstein, 1987: 3, 111). Although this view was supported by scientists in the mid-XIX Century, today «no physicist will defend it: classical mechanics has long lost its paradigmatic position» (De Regt and Dieks, 2005: 138). To stress again the diachronic difference between what was and what is conceived as understandable, Dijksterhuis (1950, 512-513) notes the fact that today anyone with a basic scientific education would judge Newton's law of inertia intelligible, while at his time it was taken as a mysterious notion.

Of course, De Regt and Dieks do not rely on the controversial variety of positions of understanding across the history, even if they pose a particular accuracy in relying on historical cases for their arguments – and we will see later on the example De Regt's uses: the gravitation debate in XVII, Thomson's model of coupled molecules, Maxwell's model of electromagnetic ether and Boltzmann's dumbbell model of diatomic gases (De Regt, 2017). So, the purpose of De Regt and Dieks is to «formulate a generally applicable, nontrivial criterion for the attainment of understanding» (2005: 38). They acknowledge a pragmatic nature of understanding and the fact

synchronically (*pace* the scientific revolutions and shifting paradigms) scientific communities are alive debating actors of what has to be taken as understanding criteria and methods. The professional disagreement – unbeatable in philosophy – is a everyday guest also in scientific practices.

that it is context-dependent. They argue that SU is «epistemically relevant and transcends the domain of individual psychology» (De Regt and Dieks, 2005: 38). The generality ambition of a philosophical account of SU is the mark of its normative attitude. While the descriptive role of cases in history of science and philosophy of science aid the scholars and the readers to find applications to conceptual frameworks to on-site scientific activities, the normative account could be helpful in disambiguating some hidden assumptions in the scientific practices and orienting the analytical work on crucial concepts and notion used in science. Also recognizing that understanding is one of the aims of science is a claim with a strong normative account, I think we can defend, if we want to foster the alliance between the sciences. Many philosophers recognize that understanding is one of the aim of sciences. We could agree with De Regt (2017: 89) that even Hempel could defend this thesis if we define understanding as the product of the covering law explanation. Moreover, also another great philosopher of the analytic tradition, Quine, in *Pursuit of Truth*, writes: «nowadays the overwhelming purposes of the science game are technology and understanding» (Quine, 1992: 20). Notwithstanding the contemporary awareness of the relevance of understanding in the science, the main doubt is whether it is possible to speak of it in general terms:

However, is it possible and useful to speak in such general terms of science and to ascribe universal aims to it? Or is the notion of understanding, if conceived as a universal aim of science, merely a blanket term, devoid of specific content? It could be argued that the idea that science has universal aims is futile and misguided, and that different scientific disciplines, and different scientist in different periods of history, simply have different aims. Indeed, a close look at the history of science reveals a wide variety of aims of scientists in different periods of history, and some historically minded philosophers have concluded from this that science does not have any universal aims at all. Laudan (1984: 138), for example, claims that it is impossible to state what the central cognitive aims and methods of science are or should be because "we have seen time and again that the aims of science vary, and quite appropriately so, from one epoch to another, from one scientific field to another, and sometimes among researchers in the same field". Two decades earlier, a similar observation was made by Toulmin (1963: 21), who argued that "the intellectual and practical activities of scientists […] have a great range and variety of purposes, which can only misleadingly be summed up in a nutshell definition (De Regt, 2017: 89).

De Regt (2005, 2017) agrees with Laudan and Toulmin that philosophers of science should take history of science very seriously and acknowledge how the aims of scientists actually vary, but he defends the idea that a general characterisation of the aims of science can be done. To account for the variety we find in science, De Regt (2005, 2017: 90) distinguishes three levels of scientific activity:

- Macro level: the science as a whole
- Meso level: scientific communities
- Micro level: individual scientists

The SU De Regt has in mind can be analysed on the macro level, considering the science as a whole with a particular aim, which on the meso and micro level can be articulated differently. The characterization of the macro level can be that «all scientist will agree that they aim to produce knowledge that is supported by experience», while the meso and micro level differences rely on the fact that «when it comes to the question of exactly how, and how strongly, scientific knowledge has to be supported by experience, the answers given by scientists from different communities, and sometimes even by scientists within the same community, will differ» (De Regt, 2017: 90). The variations in the macro level may affect the articulation of the meso and micro level. The theory of understanding proposed by De Regt focuses indeed on the micro level way of achieving SU (as individual scientists understand phenomena), and at the same time it allows for individual variation at the micro level, it is contextual and pragmatic, but general enough to be considered to have a universality aspiration. Although, the most important variations, according to De Regt, appear at the meso level:

Within a particular research community standards of intelligibility and strategies to obtain understanding are usually shared. This should not

surprise us because scientific understanding is closely related to communication: scientists who construct explanations are not merely interested in enhancing their own understanding of phenomena, but they typically want to share their insights with others. Scientific understanding is a community achievement and success in achieving understanding requires communication between scientists. When a scientist tries to communicate his or her understanding of a phenomenon to someone else (be it another scientist, a student, or a layperson), there must be shared standards of intelligibility (De Regt, 2017: 91).

The communication occurring at the meso level is the representation of the description of the profession as theoretical physicist given by Oppenheimer in the year 1948: «What we don't understand, we explain to each other» (Pines, 2015), in which the communication realizes as an explanatory dialogue between the research community. To reconcile the idea that SU is a universal, macro level, epistemic aim of science with the variations at the meso and micro level De Regt (2017, 91) proposes this distinction between the definitions of understanding phenomena (UP) and understanding theories (UT):

UP: understanding a phenomenon = having an adequate explanation of the phenomenon (relating the phenomenon to accepted items of knowledge).

UT: understanding a theory = being able to use the theory (pragmatic understanding).

Understanding as a universal epistemic aim of science De Regt has in mind is UP, it is the aim of the macro level. Moreover, De Regt (2017) argues that UT is a necessary requirement for UP. To analyse UT will be useful a detour towards the notion of intelligibility used by De Regt (2005, 2017). So far, be it enough to consider that it is sufficient to have UP in order to manage UT, while it is necessary to have UT to achieve UP. To explicate better the relation between the levels of understanding, I think it is helpful to think of it in a hierarchical terms. The mastery of a theory or a models comes for the individual scientist at the micro level; the discussion about phenomena to be explained better, to achieve SU, happens at the meso level, recalling Oppenheimer's dictum; then after sharing sometimes new standards of intelligibility an testing advanced hypothesis and explanations, the community consensus achieves scientific understanding at the macro level.

In conclusion, SU as conceived by De Regt is organized on the three level we have seen, it concerns the ways in which individual scientists and science as a whole achieve understanding, and that is the main reason to call it "scientific" understanding. He uses many examples from the history of science to elucidate this notion and defend his theory. De Regt's theory of SU is contextual in a double sense: it is sensitive to the contexts in which scientists work, may they be conceptual, physical and relational, and it is sensitive to the progress of science described by its history. Now that the main characterization of what is scientific understanding in hand, we can

study the core idea of de Regt's proposal, the intelligibility of theories.

4.3. Pragmatic Understanding through Models and Theories

The core idea of De Regt's contextual theory is that the necessary requirements for scientists to achieve SU is to have intelligible theory. De Regt recognises through the identification of UT as necessary condition for UP, a kind of epistemic superiority of theories on models and he draws a distinction between theories and models and uses the model-based understanding of scientific theories offered by Giere (2006, 59-69) to argue for the theory conception, focusing on the pragmatic dimension, still allowing for the representational success. According to De Regt, model-base explanations produce SU, but it is the theory that can successfully embed the phenomenon: «If a phenomenon is successfully embedded in a theory by means of such an explanation, we have gained scientific understanding of it (UP)» (De Regt, 2017: 32). A model-base explanation of a phenomenon is obtained by constructing a "representational" model, in Giere's terminology, that can represent the target-system so that the theoretical principle can apply to it. De Regt (2017, 32) argues for it illustrating the case of the kinetic theory of gases. Particular gas phenomena are explained on the basis of this theory and described by phenomenological laws or data models. A

phenomenological laws concerning gas phenomena is the combination of the basic experimental gas laws of Boyle, Charles and Gay-Lussac, which gives a relation between pressure (P), volume (V) and temperature (T):

$$\frac{PV}{T} = constant$$

How to reconcile this combination gas laws with the kinetic gas theory? It requires the construction of a representational gas model. The kinetic theory gives only a generic model of a gas represented as a collection of particle's in motion, under the Newton's laws of motion. To catch information about the behaviour of gas, we must add some assumptions regarding the nature of the particles and their interactions. An example of a simple representational model is the ideal-gas model. The properties of the gas particles in the ideal-gas model are described, there are smooth, hard elastic spheres, they may be conceived of as point masses, they interact only via collisions, they are (almost) not influenced by the gravity, and the collisions with the container and with other particles are elastic. These are obviously all idealizing assumptions, and with some additional statistical considerations, we can derive the relation between pressure and volume[25]. Th following formula describes the Boyle-Charles law $PV = constant$:

[25] See Feynman et al., (1963-1965).

$$PV = \frac{2}{3} N \left\langle \frac{1}{2} m v^2 \right\rangle$$

Here N is the number of particles, m is their masses, and v the velocity. In order to derive it, the macroscopic pressure is modelled as the force exerted by the particles on a unit area of the container wall. Moreover, it needs another modelling assumption regarding temperature, to derive the combined gas law. According to the kinetic theory heat is kinetic energy, and through this identification the temperature of the gas can be described as the average kinetic energy of the particles. So, we have:

$$\frac{3}{2} kT = \left\langle \frac{1}{2} m v^2 \right\rangle$$

Here k is the Boltzmann constant. Through substitution we catch the ideal-gas law, representing the phenomenological combined gas law of Boyle-Charles-Gay-Lussac:

$$PV = NkT$$

This last formula is a data model, while the former is a consequence of the representational ideal-gas model. According to De Regt's interpretation, the agreement between $PV/T = constant$ and $PV = NkT$ is an example of explanatory success of the kinetic theory. In fact, thanks to model-based explanation employing kinetic theory,

scientists first (and then also educated people) understand some basic gas phenomena.

The example given by De Regt wants to illustrate the relation between theory, models, and phenomena. Why, according to De Regt, is theory to be prior to models, and so deserving a particular treatment in developing the so-called intelligibility criteria? The priority is due to the conceptual dependency between theories and model. In the given example the theory gives the principles, i.e. laws of Newtonian mechanics and the definition of heath with kinetic energy, and the generic model of gases is constructed and articulated (to do so particular scientific skills are required, and this is one of the reason De Regt focuses on the notion of skill) to yield explanations of related phenomena. According to De Regt, the construction of a model allows the scientist to apply the theory to phenomena, in order to manipulate it and test it: «the function of a model is to represent the target system in such a way that the theory can be applied to it. In other words, models replace the bridge principles that on the traditional, Hempelian account relate a theory to empirical phenomena» (De Regt, 2017: 34). According to Morgan and Morrison (1999) the models "mediate" between theory and phenomena and Cartwright (1983) offers a detailed description of this connection through her "simulacrum" account of explanation: «to explain a phenomenon is to construct a model that fits the phenomenon into a theory» (Cartwright, 1983: 17). De Regt is on the same line of thought, and he uses this description to stress the

pragmatic requirements for scientist to construct model, according to which – De Regt says – understanding is an epistemic skill of scientists. He is right in claiming that without skills (specific skills in scientific areas), science cannot be done. Nonetheless, I think it is wrong to deflate the understanding to the pragmatic requirement of possessing skills. It is not the only condition in play. I submit that to have SU scientists must hold a representational link to the phenomena under scrutiny, in virtue of which they can distinguish correct from wrong inferences about such phenomena, being the inferences true or false according to the contexts in which science is done. Without skills, scientists cannot perform their job, but without representational accuracy (which is bound with knowledge and truth), they cannot achieve understanding. Arguing for the pragmatic requirement, he affirms:

> In the modelling stage, the target system is presented in such a way that the theory can be applied to it: we decide to describe system *S* as if it is an *M* (where *M* is a model of which the behaviour is governed by the principles of the theory). The construction of models is not a matter of deduction but a complex process involving approximations and idealizations. There are no algorithms of formal principles that tell us how to get from the description of a real system to a suitable model. […] This implies that the construction of model, and accordingly the construction of an explanation, is a process in which scientists have to make pragmatic decisions and must accordingly rely on skills and judgments (De Regt, 2017: 34).

This view about the pragmatic understanding of phenomena relies on the idea that model-building necessarily involves idealization and approximation, which accordingly require skills and judgment. De Regt recalls that a defence of this thesis was proposed by Putnam (1978: 72), who explicitly identifies the «unformalized practical knowledge» with skills. De Regt points out that Putnam's analysis concerns the context of discovery – the construction of the models - and also the context of justification – the ability to assess whether a model is good or not. Pragmatic decisions are involved in the science work, and De Regt claims that SU is a matter of skills, it is namely pragmatic understanding. Now, with this in mind, let us see what he intends as pragmatic understanding and what role intelligibility of theories plays in it.

4.4. Intelligibility

Going from models to theory, and vice versa, scientists need specific skills. These are enough according to De Regt to have SU, namely pragmatic understanding. What are the nature and conditions of pragmatic understanding?

Scientist seek explanations that fit the phenomenon to be explained into a theoretical framework and connect it with relevant background knowledge. The connection between the phenomenon and theoretical and background

knowledge is typically made through models. The construction and evaluation of such model-based explanations involves making suitable idealizations and approximations, and this require pragmatic understanding: scientists need to make the right judgments regarding idealization and approximation, and possess the right skills to build a model on this basis (De Regt, 2017: 36).

Right skills are connected to theoretical virtues of simplicity, visualizability, but not all scientists value the same qualities in the same way. De Regt does not take a specific stance in the debate about the theoretical qualities that play a crucial role in construction and evaluation of scientific theories. So, he proposes this definition of intelligibility:

> *Intelligibility*: the value that scientists attribute to the cluster of qualities of a theory (in one or more of its representations) that facilitates the use of the theory.

It means that if a scientist understands a theory, the theory is intelligible to her. In this definition intelligibility is not an intrinsic property of theories, but an extrinsic and relational property. In fact, it depends not only on the qualities of the theory, but also on the skills of the scientists who work with it. Intelligibility is also a contextual measure of the fruitfulness of a theory (De Regt, 2017: 40); a theory can be fruitful for scientists in one context, but not so useful for scientists in other contexts. For example, in 1926 the wave

mechanics advanced by Schrödinger gained more popularity among the physics community than the matrix mechanics developed by Heisenberg, due to the fact that wave mechanics was visualizable (more and easier intelligible) and mathematically less intricate[26]. Schrödinger wave mechanics was «successfully applied to a great variety of problems unamenable to matrix treatment» (Beller, 1999: 36). To the general physicists community in the 1920s wave mechanics was more intelligible than matrix mechanics. According to De Regt (2017) and Beller (1999) it was its theoretical qualities, namely its visualizability and mathematical simplicity, that allowed scientists to construct models based on it to explain certain phenomena. On the other hand, proponents of the matrix mechanics do not consider visualizability as a necessary condition for intelligibility; in fact, Heisenberg was able to develop the foundation of matric mechanics without looking at a visualizable model of atomic structure[27]. But from the intelligibility of a theory derives the ability to use it to different purposes: «Scientists prefer a more intelligible theory to less intelligible one, not because it gives them the feeling of understanding but rather because they have to be able to use the theory in practice» (De Regt, 2017: 41).

[26] We will see that the same process applies to the case of Bjorken scaling, described by Khalifa (2017), which I will discuss in Chapter 5. Also in this case, we assist to an improvement of the explanatory information core given by the partons' model advanced by Feynman, which refines the explanatory structure of Bjorken's argumentation.
[27] See De Regt (2017, Ch. 7) for a detailed analysis of this case.

4.5. A Contextual Theory of Scientific Understanding

De Regt's dialectics for the development of his contextual theory of scientific understanding (CTSU) can be described in the following structure of premises and conclusion:

The aims of science require intelligible theories (I):
P1) providing correct explanations is an epistemic aim of science;
P2) correct explanations require scientists to use intelligible theories;
C) the epistemic aims of science require intelligible theories.

SU requires the use of intelligible theories, developed, discussed, described by scientists. The premises of his argument are that one of the aims of science is to provide correct explanations and that correct explanations require scientists to use intelligible theories. He does not offer a definition of empirical adequacy and consistency, but he asserts that they are research values because «there may be variation in how these values are ranked and applied in specific cases» (De Regt 2017: 38). The theoretical quality of SU and the adequacy and consistency of the explanations SU is provided with are the main cores of De Regt's framework.

- *UP*: understanding a phenomenon = having an adequate explanation of the phenomenon (relating the phenomenon to accepted items of knowledge).

- *UT*: understanding a theory = being able to use the theory (pragmatic understanding).

According to De Regt (2017) UP necessarily requires UT, which was articulated with the help of the pragmatic notion of intelligibility. Furthermore, he distinguishes between the criterium for understanding a phenomenon, related to UP, and the criterion for the intelligibility of theories, related to UT.

The Criterion of Understanding Phenomenon (CUP) is so defined:

- *CUP*: A phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency.

The basic idea of the theory is the thesis that explanatory understanding of phenomena requires intelligible theories, where intelligibility has been defined as the value that scientists attribute to the cluster of qualities that facilitate the use of a theory. (De Regt, 2017, p. 88)

CUP has two features:

1) First, in keeping with his contextualism, de Regt uses examples from the history of science to argue that scientific explanations do not conform to a single template.

2) Second, de Regt construes empirical adequacy and consistency as values because "there may be variation in how these values are ranked and applied in specific cases" (2017, p. 38). He takes them to be "basic" because, unlike the aforementioned values (simplicity, scope, familiarity, causation, and so on), every scientific explanation must exhibit these two values to some degree.

The utilization of historical examples from the annals of science underscores de Regt's rejection of a singular template for scientific explanations. By drawing on a variety of cases from the history of science, de Regt demonstrates the diversity and flexibility inherent in scientific explanation. This contextualist perspective acknowledges that explanations can vary widely depending on the scientific context, challenging any rigid or universal model of explanation.

De Regt treats empirical adequacy and consistency as fundamental values within his framework. Unlike other values such as simplicity or scope, which may vary in their application across different scientific contexts, empirical adequacy and consistency are deemed essential in every scientific explanation. De Regt's rationale for this stems from the notion that these values serve as foundational

criteria for evaluating the credibility and reliability of explanations. By considering empirical adequacy and consistency as basic values, de Regt emphasizes their universal significance in the epistemic landscape of scientific understanding.

Overall, these two features of CUP reflect de Regt's nuanced approach to understanding scientific explanation, which emphasizes the importance of historical context and the fundamental role of empirical adequacy and consistency in evaluating scientific claims.

Moreover, since the contextual features of his theory, also the property of intelligibility has to be characterized as context-sensitive. Thus, he proposes the Criterion for the intelligibility of theories, CIT_1, that has to be consistent with his claim:

- *CIT_1*: a scientific theory *T* (in one or more of its representations) is intelligible for scientists (in context *C*), if they can recognize qualitatively characteristic consequences of *T* without performing exact calculations.

De Regt argues that different scientists place greater value on different clusters of qualities. Scientists' skill are the crucial contextual determinants of which of these clusters furnish intelligibility.

De Regt's criteria for scientific understanding are concise and comprehensive, emphasizing the importance of theoretical intelligibility and adherence to fundamental epistemic values.

According to these criteria, scientific understanding of a phenomenon *P* requires an explanation based on an intelligible theory *T*, while also meeting the basic requirements of empirical adequacy and internal consistency. This framework ensures that explanations are grounded in coherent theoretical frameworks and supported by empirical evidence. Additionally, De Regt introduces CIT_1, which defines the intelligibility of a scientific theory within a specific context. CIT_1 enables scientists to qualitatively recognize characteristic consequences of *T* without exact calculations, providing a practical measure for assessing the intelligibility of scientific theories. By incorporating these criteria, De Regt's framework offers a structured approach to evaluating scientific understanding, ensuring that it is robust, grounded, and aligned with core epistemic values.

4.6. Visualizability, Intelligibility and the Quantum Mechanics

The reflection on scientific understanding made by de Regt (2017), origins from the link between intelligibility and visualizability[28] pointed out by Schrödinger, that, on the basis of his conviction that nature is understandable (Schrödinger [1954] 1996), argued in favour of the visualizability of theories in space and time to

[28] For a detailed analysis of visualizability (*Anschaulichkeit*) as conceived by Schrödinger, see De Regt (1997).

obtain understanding of nature. Indeed, «we cannot really alter our manner of thinking in space and time, and what we cannot understand (*verstehen*) within it we cannot comprehend at all. There *are* such things–but I do not believe that atomic structure is one of them» (Schrödinger 1928, 27). Chapter 7 of De Regt's *Understanding Scientific Understanding* delves into the challenges of visualizing and grasping the intelligibility of quantum theory, using examples primarily from Schrödinger and Heisenberg. De Regt explores how quantum mechanics (QM), with its abstract mathematical formalism and non-intuitive concepts, presents difficulties in providing a visualizable representation of its underlying processes. He discusses Schrödinger's development of the wave function, which was initially intended to represent real waves, offering a seemingly intuitive visualization of quantum systems. However, De Regt highlights the limitations of this approach, noting that the wave function's interpretation as a probability amplitude deviates from classical wave behaviour. He also delves into Heisenberg's matrix mechanics, which introduced an entirely abstract formalism based on matrices, challenging traditional notions of visualizability in physics. De Regt analyses how these differing approaches to visualization reflect broader debates within the philosophy of science about the role of visualization in scientific understanding. He suggests that Schrödinger's proposal gained popularity thanks to its intelligibility property, but as Rovelli

(2021)[29] highlights it is through a technical and conceptual step that he made to visualize better the main features of QM: «The technical step was to translate the unfamiliar algebraic language of quantum theory into a familiar one: differential equations. This brought the novel ethereal quantum theory down to the level of the average theoretical physicist» (Rovelli 2021: 1). This is what made him win, so to say, the visualizability prize of QM in the 1920s. The conceptual step, instead, «was to introduce the notion of "wave function", which soon evolved into the general notion of "quantum state", Ψ, endowing it with ontological weight» (*Ibidem*). In his analysis, de Regt (2017) emphasizes that while visualization can aid understanding to some extent, it cannot fully capture the complexities of quantum phenomena. Ultimately, De Regt's exploration underscores the tension between the desire for visualizable representations in science and the need to grapple with abstract formalisms to achieve true scientific understanding in the realm of quantum mechanics.

[29] Rovelli (2021) criticizes the intelligibility Schrödinger achieved in the representation of the wave function in his defended Relational Interpretation of Quantum Mechanics: «The mistaken idea is that the quantum state ψ represents the "actual stuff" described by quantum mechanics. This idea has pervaded later thinking about the theory, fostered by the toxic habit of introducing students to quantum theory in the form of Schrödinger's "wave mechanics", thus betraying history, logic, and reasonableness» (Rovelli 2021: 1). This point was suggested by Fano.

4.7. Objections to the Contextual Theory of Scientific Understanding

In this conclusive paragraph of this chapter, I have detailed five objections to the contextual theory of scientific understanding proposed by de Regt.

1 – The first objection concerns the argument for his first central thesis, that science requires intelligible theories. Its first premise (P1 above) states that correct explanations are among science's epistemic aims. De Regt is not explicit on what it means for something to be an epistemic aim. Consider a fairly standard empiricist gloss on explanatory value: explanations are of epistemic value only insofar as they are a means to saving the phenomena. On such a view, explanation is not an epistemic aim of science because its epistemic value is wholly instrumental. Thus, any non-instrumental explanatory value is non-epistemic, e.g. explanations' utility as psychological crutches or in serving scientists' practical interests. Absent further discussion of his axiological framework, it is unclear how de Regt distances himself from this particular conception of explanation and understanding. However, he cannot accept this conception without rendering the aforementioned argument unsound.

2 – One may challenge De Regt's contention that correct explanations need only be intelligible, empirically adequate, and

consistent. For instance, just-so stories appear capable of meeting these criteria, but it is hard to see why such ad hoc explanations would confer understanding or serve as suitable epistemic aims of science. De Regt's treatment of empirical adequacy and consistency as values (rather than as strict requirements) makes these constraints even more malleable, which further compounds this worry. Had De Regt engaged these problems head-on, these concerns might well have been assuaged.

3 – The criterion CUP requires explanations to be "based on" intelligible theories. It is unclear what this means. Given the pains to which de Regt goes to establish that understanding is not merely subjective, an explanation that is nothing more than psychologically associated with certain theoretical elements does not seem "based on" that theory. However, De Regt only tells us that this basing relation is not restricted to deriving the explanation from the theory. Similarly, it is unclear how much theory understanding requires. On the one hand, to defend CUP's sufficiency, De Regt wishes to rule out intelligent design as an intelligible theory, on the grounds that it lacks theoretical principles that "can be used to construct specific, empirically adequate explanations of concrete phenomena" (p. 95). On the other hand, to defend CUP's necessity, De Regt claims that several special sciences -- most notably psychology and sociology -- only rely upon "loosely circumscribed theoretical principles" (p. 97). However, it is not obvious that these latter theoretical principles are

in better standing than intelligent design's when it comes to constructing good explanations

4 – The connections between CUP and more traditional philosophical accounts of scientific explanation are occasionally puzzling. It would be in the spirit of De Regt's contextualist approach if he first identified an explanation using whatever theory of explanation was apt for the phenomenon, and then used criteria such as CIT1 to determine whether or not the explanation provided understanding of that phenomenon. However, at times, de Regt seems eager to show that CUP not only tells us when an explanation provides understanding, but when it is an explanation at all. For instance, he feels obliged to show that CUP "solves" some well-known puzzles that any account of explanation ought to address, such as the fact that the length of a shadow does not explain the height of the flagpole that casts it. His response is essentially a bullet-biting one: in some contexts, shadows explain heights. However, the burden of an adequate solution to the symmetry problem would only be a worry if CUP had the further requirement that an explanation of P is based on an intelligible theory T if a description of P is a consequence of some of T's principles. Since CUP does not entail this, and De Regt seems reluctant to characterize how explanations are "based on" theories in purely deductive terms, it is unclear why he has shouldered this particular burden of proof (Khalifa, 2017).

5 – Khalifa about De Regt's definition of SU as distinct from knowledge in that requires specific skills. Khalifa's critique of De Regt's definition of Scientific Understanding (SU) as distinct from knowledge highlights the requirement for specific skills. Khalifa questions the delineation between understanding and knowledge in De Regt's framework, particularly concerning the acquisition and application of specialized skills. This critique prompts further examination of the relationship between understanding, knowledge, and expertise within the context of scientific inquiry.

Chapter 5:
Khalifa's Account of Scientific Understanding

In this chapter, we explore Khalifa's nuanced account of scientific understanding, which offers valuable insights into the epistemic aims and methodologies of scientific inquiry. Khalifa's framework emphasizes the distinction between descriptive and explanatory scientific understanding, highlighting the role of both empirical evidence and theoretical coherence in fostering comprehension of natural phenomena. Central to Khalifa's approach is the recognition that scientific understanding entails more than mere factual knowledge; it involves the ability to construct coherent explanations grounded in empirically adequate theories. Through an analysis of Khalifa's work, we delve into the complex interplay between empirical data, theoretical frameworks, and the cognitive processes underlying scientific reasoning. Furthermore, Khalifa's emphasis on the context-dependence of scientific understanding sheds light on the

diverse methodologies employed across different scientific disciplines. By examining Khalifa's account, we gain a deeper appreciation for the multifaceted nature of scientific understanding and the nuanced epistemic practices that underpin scientific inquiry.

5.1. Introduction

Enriched by wit and creativity, Khalifa's work is full of new labels and sharp epistemological thinking about understanding. His view is fully described in *Understanding, Explanation, and Scientific Knowledge* (Khalifa, 2017), a title that already advances the main thesis of the book: from the philosophy of understanding we can derive claims to encapsulate an account of SU, which is deeply bound with the notion of explanation and scientific knowledge. From this starting point, it is clear that Khalifa's view is rooted in the tradition of the received view about understanding. His position is the main opponent of De Regt's account we can fin in literature nowadays. Nevertheless, maybe "opponent" is not the right term to be used. In fact, Khalifa and De Regt have recently engaged themselves in a discussion opening to a third view on SU, which can be better solve some of the problems the contextual theory and Khalifa's view have. As De Regt, also Khalifa recognizes the understanding can be «understood in different ways» (Khalifa, 2017: 1). It is the theme of the variety of standards of understanding

already discussed by De Regt; of course we can find not only divergence in scientists opinion about how to conceive understanding and understandability standards, but there is also a theoretical variety, namely philosophical, about how to understand understanding and scientific understanding. On the other hand, Khalifa focuses on explanatory understanding. He argues for the reduction of objectual understanding to explanatory understanding.

5.2. Khalifa's EKS Model of Understanding

It is a common experience to acknowledge different degrees of understanding concerning a phenomenon. In Chapter 2. we have seen the example of EKS model applied to the anthropological study of the Ilongot tribe. Khalifa's defence rests on his Explanation-Knowledge-Science (EKS) model of understanding, which is underpinned by four key ideas. Firstly, understanding is not binary but exists on a spectrum, allowing for comparisons of understanding levels among individuals regarding a specific phenomenon. Secondly, the depth of one's understanding correlates with how thoroughly they grasp the explanatory nexus concerning the phenomenon, defined as the collection of correct explanations and the relationships between them. Thirdly, grasping entails achieving a cognitive state resembling scientific knowledge about the explanatory nexus, implying that superior understanding entails a

closer alignment with scientific understanding. Lastly, to safeguard against coincidental similarities, Khalifa introduces the requirement of scientific explanatory evaluation (SEE). This involves considering multiple plausible explanations, comparing them using rigorous scientific methods and evidence, and aligning one's beliefs accordingly. These principles offer criteria for assessing the resemblance between an individual's understanding and scientific knowledge, including the consideration of alternative explanations, the rigor of comparative analysis, the scientific validity of criteria employed, the reliability and accuracy of beliefs, and the versatility of utilizing explanatory information to attain scientific objectives

Since SU comes with degrees, an agent can understand *p* better then someone else. To account for the better understanding of *p*, Khalifa (2017: 4) uses the comparative principles:

- *Schema for Comparative Understanding (SCU)*: *ceteris paribus*, S_1 understands why *p* better than S_2 if and only if S_1 has minimal understanding of why $p + X$.

This schema assumes that even if we recognize that Alice understands why the sky is blue better than Bob, there is a minimal understanding as a condition for having better understanding. It is that if Alice has better understanding of *p* than Bob, Alice has at least a minimal understanding of *p*, while we cannot say the same for Bob. The SCU holds that Bob could also not retain any understanding of

p. As for a definition of minimal understanding (MU), Khalifa argues that:

- *Minimal Understanding (MU)*: *S* has minimal understanding of why *p* if and only if, for some *q*, *S* believe that *q explains why p*, and *q explains why p* is approximately true.

According to Khalifa (2017), we can take MU as the cornerstone of understanding. A question could be whether also scientific understanding implies a switch between minimal and maximal scientific understanding. On the opposite to MU, there is the maximal understanding, that Khalifa labels ideal understanding:

- *Ideal Understanding (IU)*: *S* ideally understands why *p* if and only if it is impossible for anyone to understand why *p* better than *S*.

This ideal understanding is a sort of ideal limit of understanding. But, in the case of scientific understanding, since SU is contextual and changes according to the scientific knowledge, methodologies and tools at disposal, also IU turns out to be contextual.

The notion of ideal understanding is related to the notion of outright understanding, that Khalifa develops using a contextualist semantics. He conceives the outright understanding as a non-comparative principle:

- *Outright Understanding (OutU)*: "*S* understands why *p*" is true in context *C* if and only if *S* has minimal understanding and *S* approximates ideal understanding of why *p* closely in *C*.

The contextuality of understanding is conserved also in depicting the account of generic understanding, to describe the kind of understanding a person has of an object of phenomenon, but without a deep grasp of its functions, organization and features:

- *Generic Understanding (GU)*: *S* has some understanding of why *p* if and only if "*S* understands why *p*" is true in some context *C*.

As we see, Khalifa (2017) wants to depict a broad account of understanding that can give the epistemological basis for the definition of an account of scientific understanding. For the construction of the SU account, Khalifa needs two comparative principles: nexus principle and the scientific knowledge principle, we will see in the next paragraphs.

5.3. The Nexus Principle

The nexus principle main core is the following claim: «a natural suggestion is that explanatory understanding is the possession or "grasp" of an explanation» (Khalifa, 2017: 6). For example, to understand why the rainbow is colourful is to have a correct explanation of why the rainbow is colourful. Multiple factors contribute to the richness of colours of the rainbow. It happens when light enter a drop of water, gets refracted and reflected back into the observer's eyes. Depending on the media involved, light runs at different velocities. It travels faster through air, even faster outside the atmosphere (it is 1.0003 times slower in the atmosphere on earth than in a vacuum; it slows down from 299.792,458 meters per second to 299.702,547, a difference of 89,911 meters per second), it travels slower through water. As the light enters a raindrop, it happens after the rain, when droplets of rain are still moving in the air, it slows down, so it bends a little and gets refracted. Not all frequencies of light get refracted at the same angle. In fact, the colours with shorter wavelengths, i.e. blue, indigo, violet, get refracted less than the longer wavelengths, i.e. red. This phenomenon spreads colours out as they were divided through a prism. The light gets reflected off pointing at the back of the droplet, then it get refracted another time as it goes out through the raindrop and travels to the observer's eye. The rainbow appears to be circular due to the specific range of

angles, 40 (violet) to 42 degrees (red), the light gets refracted and reflected to the observer's eyes.

The nexus principles: ones understanding of the coloration of the rainbow increases as one gathers more of the correct explanatory information, and also learns more about how the information hang together. Khalifa is very careful in linking the increasing of the explanatory information at disposal with the learning of how the information are related together. He defines this amalgama as the explanatory nexus of *p* (be *p* the formation of a rainbow): «let the *explanatory nexus* of *p* be the set of correct explanations of *p* as well as the relations between those explanations». The *Nexus Principle* (NP) is so formed:

The Nexus Principle: *ceteris paribus*, if S1 grasps p's explanatory nexus more completely than S2, then S1 understands why p better than S2.

NP makes appeal to the ability of an agent to "grasp" explanatory information about a phenomenon. Khalifa should distinguish between grasping and understanding.

5.4. The Scientific Knowledge Principle

Khalifa's Scientific Knowledge Principle (SKP) provides a structured approach to evaluating understanding based on the degree to which an individual's comprehension aligns with scientific knowledge. This principle recognizes that understanding is not absolute but relative, allowing for comparisons between different individuals' grasp of a phenomenon. By emphasizing the resemblance between an individual's understanding and scientific knowledge, Khalifa underscores the importance of aligning one's comprehension with established scientific frameworks. This principle encourages a rigorous examination of explanatory nexuses, promoting the adoption of scientific methods and evidence in evaluating alternative explanations. By prioritizing the convergence with scientific knowledge, the principle facilitates a more robust and reliable assessment of understanding, fostering a deeper engagement with scientific concepts and phenomena.

The Scientific Knowledge Principle (SKP): *ceteris paribus*, if S1'2 grasp of *p*'s explanatory nexus bears greater resemblance to scientific knowledge than S2's, then S1 understands why p better than S2.

According to Khalifa, SKP relates to the difference of understanding between an atmospheric physicist about the features of the rainbow

and her students. He also recognizes that the greater quality of the physicist's understanding in comparison to her students it is not just a matter of quantity of explanatory information at disposal, but also a matter of skill, mastery of background theoretical knowledge, methods and evidence. SU is firmly connected to expertise (Wilkenfeld, Plunkett and Lombrozo, 2016) and skills (De Regt, 2017). According to Khalifa, SKP can account for the requirement of a scientist to have specific skills, since he focuses on empirical phenomena and scientists are the leading experts on the issue. To strengthen SKP, then, Khalifa specifies what is the scientific knowledge of an explanation and how should we understand «a grasp's resemblance» (Khalifa, 2017: 11) to scientific knowledge. From scientific knowledge we systematize explanations, from explanations we achieve understanding. If this inference is retained correctly, one may genuinely ask: When does a subject have scientific knowledge (SK)? Khalifa, then, advances this definition of SK:

Scientific Knowledge (SK): *S* has scientific knowledge that *q explains why p* if and only if the safety of *S*'s belief that *q explains why p* is because of her scientific explanatory evaluation.

According to this definition, a specification of safety is needed. Khalifa follows the depiction of safety given by Pritchard (2005), Sainsbury (1997), Sosa (1999), and Williamson (2000); it is related

to the property of the beliefs a person has, in particular, according to their definitions, a belief is safe in case the belief-forming process of the believer could not have easily led to a false belief. It is a kind of falsity-filter belief forming process. Khalifa translates this principle in the case of SU with what he calls *scientific explanatory evaluation*, that is the scientific evaluation (the scientific control over the safety of the scientific beliefs) of an explanation. He submit that the scientific evaluation of an explanation (SEEing) has three features, consideration, comparison and belief-formation:

- *Consideration*: scientists consider many plausible potential explanations of phenomena under scrutiny. Correct explanations are the ones that for realists succeed to be approximately true, and for antirealists satisfy empirical adequacy, as a more modest requirement (even if I submit that an explanation to satisfy the empirical adequacy criteria has to be approximately true).
- *Comparison*: scientists have the ability to compare the potential explanations that have been considered. They use the best methods, evidence, and also non-evidential considerations of the area in which they work, may it be natural or social sciences. Another difference between Khalifa's account and De Regt's one, is that the contextual theory focuses mainly on natural science, even if De Regt assumes its generality to be enough broad to accommodate

also the requirements of social science, while the SKP is explicitly formulated to account also for social sciences.

- *Belief-formation*: scientists form beliefs – doxastic attitudes – based on the comparison discussed: «Scientists believe that clear winners in the prior stage of comparison are correct, believe that clear losers are incorrect, and assign appropriate degrees of belief about the middle of the pack» (Khalifa, 2017: 13). On the basis of the safety property of belief-formation, also the scientific doxastic activity is guided by safety protocols, according to Khalifa.

To conclude, SEEing is centred on the notion of safety. Based on these comparisons scientists form doxastic attitudes, it entails that scientific knowledge require safe belief. Khalifa in particular draws SEEing having in mind this notion of safety (Pritchard, 2009: 34): «S's belief is safe iff in most near-by possible worlds in which S continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which S continues to form her belief about the target proposition in the same way as the actual world, her belief continues to be true». In this way, SEEing secure the correctness of the scientific evaluation of explanations and warrants the degree of scientific understanding to which the agents refer to.

5.5. The Explanation-Knowledge-Science (EKS) Model

The core of Khalifa's account of SU is the Explanation-Knowledge-Science model (EKS), that is described below. To have a satisfactory account of what does it mean to have better understanding, we need SKP, which can be formulated as the Explanation-Knowledge-Science (EKS1) principle:

EKS1: S_1 understands why *p* better than S_2, if and only if:

A) *ceteris paribus*, S_1 grasps p's explanatory nexus more completely than S_2; or

B) *ceteris paribus*, S_1's grasp of p's explanatory nexus bears greater resemblance to scientific knowledge than S_2's.

When an agent has minimal understanding of a phenomenon, instead, Khalifa submits the following definition of the second principle (EKS2):

EKS2: *S* has minimal understanding of why *p* if and only if, for some *q*, *S* believes that *q explains why p*, and *q explains why p* is approximately true.

Khalifa call these claims together, since they rely on explanation, knowledge and science as key-notions, the EKS Model of Understanding. EKS1 and EKS2 are sufficient conditions for

providing accounts of ideal, outright and generic understanding. Khalifa uses the notion of outright understanding in the way De Regt uses the meso-level dimension of understanding, it is to conceive SU and everyday understanding on a single continuum (Khalifa, 2023: 15). The picture of the spectrum of understanding is then as follows: from minimal understanding to everyday understanding, to scientific understanding, to ideal understanding.

What is essential in terms of having scientific understanding is the ability to compare difference potential explanations that have been considered: scientists cite evidence and non-evidential scientific factors: simplicity, scope, mechanism, conservatism, analogy, testability, unification, fruitfulness and other theoretical virtues that figure in scientific activity. Sometimes one explanation sorts out as the winner of the comparison, while it can happen that multiple explanations are good along different dimensions[30]. The interplay between explanations, knowledge and science is the hinge that provides scientific understanding of phenomena, as describe by EKS model, as an upgraded version of the received view.

5.6. The Case of Bjorken Scaling

The example chosen by Khalifa to show the application of the EKS model comes from the history of high energy physics

[30] It is the case of explanatory integration I will discuss in Chapter 6.

experiments, and regards mainly the physicists James Bjorken and Richard Feynman, and their research teams. In particular, Khalifa tests the plausibility of EKS Model applying it to the case of Bjorken scaling. In the 1960s, Bjorken (Professor Emeritus at the SLAC Theory Group at the Stanford Linear Accelerator Center, and winner of the Dirac Medal of the ICTP in 2004, the Wolf Price in Physics and EPS High Energy and Particle Physics Prize in 2015) made a novel prediction about a specific kind of scaling.

He used abstract tools opaque to ta majority of physicists at that time, even to experimental physicists who confirmed Bjorken's prediction. In 1968 he discovered the light-cone scaling, that is a phenomenon in the deep inelastic scattering of light occurring on strongly interacting particles, such as protons and neutrons, also known as hadrons. When they are under high energies, hadrons behave like virtually independent point-like constituents. These hadrons have scaling properties: they are determined by dimensionless kinematic quantities, i.e. scattering angle or the ratio of the energy to a momentum transfer. Since increasing energy implies spatial resolution, scaling entails the independence of the absolute resolution scale, and so point-like substructure. Later on, the concept of scaling has been reformulated by Feynman in the parton model, which has been used to understand quark composition of hadrons at high energy. Feynman gave Bjorken complex mathematical model a physical interpretation that was intelligible to

a much wider audience (remind the De Regt's intelligibility story from Heisenberg to Schrödinger).

As we have seen, Khalifa applies his EKS Model to the case of Bjorken's scaling, or scaling invariance. In the 1960s a team of scientists from the Stanford Linear Accelerator (SLAC) and the Massachusetts Institute of Technology (MIT) produced many experimental measurements of the scatter resulting from firing electrons at a proton target. They want to measure the cross-section σ, which is the likelihood of an interaction between particles. The aim was to discover the properties of subatomic particles. In particular at that time hadrons were ones of the most studied subatomic particles, which are affected by nuclear strong forces, while leptons are supposed to be unaffected by such forces. According to hadrons-leptons distinction, neutrons, protons, kaons and pions are hadrons, while electrons, muons and neutrinos are leptons.

The SLAC-MIT team examined in the 1967 two kind of scattering phenomena: elastic scattering in which electrons and protons-target retain their identities; and inelastic scattering in which the interactions between electrons and protons-target do not make the protons retain their identity. The background knowledge established before the experiment, based on the quantum electrodynamics (QED), was that the cross-sections for elastic and inelastic scattering would decrease quickly when electrons beams were fired at higher energy levels, and scatter was measured at larger angles. According

to QED, electrons should interact as hard, point-like entities, while protons were conceived as having a diffuse, soft structure with a finite volume of space. It means that, if QED theory was correct, the scatter at high energies and large angles would be very little, because the soft structure of protons would permit electrons to strike glancing blows. Since the 1967 the experimental results were consistent with the QED, but the SLAC-MIT team used higher energy than their predecessors and they found a surprising exception. Almost all their results were consistent with the prior theory, but the surprising exception concerned the discovery (in contrast with QED predictions) that the cross-sections for the interactions between electron-proton and electron-electron are roughly the same as inelastic scattering at high energies and large angles. The discovery shows that the electron-proton interactions have higher cross-sections than was expected for deep inelastic scattering, according to QED predictions. In conclusion, the discovery suggests that the proton is composed of hard point-like entities and not as soft as the QED suggested it to have. Bjorken was not surprised by the result, he had already in 1966 predicted the proton "hard point-like entities" composition using an esoteric mathematical framework for that time in quantum field theory, the current algebra (Bjorken, 1967). He predicted that «the absolute energy of an experiment does not determine the cross-section of electron-proton scattering, which is consistent with the SLAC-MIT team's surprising result that these cross-sections do not decrease at higher energy levels», and

moreover, Bjorken claims that the cross-section of deep inelastic scattering, σ_{DIS}, is determined by the «ratio of the energy loss of the scattering electrons *v* to the momentum transfer between the electron and the proton *q*» (Khalifa, 2017: 29). In April 1968 he plot the two functions W_1 and W_2, representing the proton's structure against the variable $\omega = -q^2/Mv$, which is now known as the Bjorken scaling variable, where *M* is the proton's mass. He then predicted that the results concerned unique curves, Bjorken scaling curves.

Khalifa tells the story of Bjorken's scaling exploiting two main *explananda*, recalling that scientific knowledge of why something is the case, what is to be explained, the *explanandum*, requires former knowledge that it is the case. The first explanandum is the scattering phenomenon (SP):

Scattering phenomenon: Why is $\frac{\sigma DIS}{\sigma MOTT} \approx 1$ (rather than < 1).

Here, σ_{DIS} is the cross-section of deep inelastic scattering, and σ_{MOTT} is the cross-section of electron-electron scattering. The parenthetical contrast describes the predicted result of QED. The SLAC-MIT experiment shows that, in contrast to QED, being protons not so soft as they were supposed to be, electron-proton scattering is more similar to electron-electron scattering than the prior theory predicts. The second explanandum regards Bjorken scaling laws. A scaling law is a function f such that $f(cx) \propto f(x)$, where *c* is a constant. So, the shape of the function is preserved when the size or the scale of

the function's argument changes. This is what happens with the proportional feature (scaling) of the equation expressing the area A of a square as a function of the length l of one of its sides: $A(l) = l^2$. Bjorken identifies a relation parameter that holds regardless of the quantities involved; there are two Bjorken scaling laws (BSL), which are our second explananda:

Bjorken scaling laws: $W_1 = F_1(\omega)$; and $\nu W_2 = F_2(\omega)$.

Since the proportion pattern holds, W_1 scales like $F_1(\omega)$, and accordingly W_2 scales like $F_2(\omega)$. The scientists were interested in understanding why these scaling relations hold, i.e. why an increase in ν implies a necessary proportional increase in F_2. We can accept that the scientists at that time could take SP and BSL as true, and we can say that they *knew* that the explananda were true. Here it is the difference between knowing that something is the case, in fact we can know that an explanandum is the case and that is true, but at the same time we lack understanding of why it is true. With the help of Khalifa's Nexus Principle, concerning this example, we can take that SU «advances when we start to grasp correct explanations of these two phenomena» (Khalifa, 2017: 31). Moreover, Khalifa notes that Bjorken's account of scaling is a sketch of an asymptotic explanation[31], which requires the identification and elimination of

[31] See Betterman (2002) for a philosophical analysis of asymptotic explanations.

irrelevant behaviour in question. This is the general schema for a kind of asymptotic explanation (AES), given by Khalifa (2017: 31):

Asymptotic Explanation Schema:

Explanation Target: Why does the same pattern of behaviour emerge in diverse physical systems?
Explanatory Pattern: The pattern of behaviour can be expressed as a mathematical function.
Various details about these systems are constant in the asymptotic limit of this function.
Given the underlying microphysics of the system, differences in these details are irrelevant to the pattern of behaviour.

If we take a function to an asymptotic limit, we identify the behaviour of the function when one or more of this arguments approaches to zero or to infinity. In the AES, we can identify the goal of asymptotic explanations as to provide a mathematical representation of the pattern of behaviour emerging in the limit. In asymptotic reasoning it is used dimensional analysis, which involves «examination and manipulation of the various dimensions involved in a given problem, with the goal of creating dimensionless parameters» (Khalifa, 2017: 32). Following these lines, Bjorken gives an asymptotic explanation of BSL, using dimensional analysis, and in his account of inelastic scattering, he relies on effective mass,

which is a dimensional constant. Since dimensionless quantity are useful in cases in which asymptotics are in play, he created the dimensionless Bjorken scaling variable:

$$\omega = -q^2 / Mv$$

Here v is the energy lost by the colliding electron, and q^2 is the square of the momentum transfer between the electron and the proton; they are both taken to infinity, but the ratio holds constant $\omega = v/q^2$. The following are the resulting limits:

$$\lim_{q^2\to\infty}\left(\frac{v}{q}\right) fixed \quad vW_2(q^2, v) = F_2\left(-\frac{q^2}{Mv}\right) = F_2(\omega); \text{ and}$$

$$\lim_{q^2\to\infty}\left(\frac{v}{q}\right) fixed \quad MW_1(q^2, v) = F_2\left(-\frac{q^2}{Mv}\right) = F_1(\omega).$$

Bjorken with these two limits represents the fact that the determination of the values of the structure functions relates to the dependency on each other of q^2 and v. This point can be observed only the values approach extreme conditions (∞); when $q^2 \to \infty$, they are at very high energies. The meaning of the other part of the equation represents that W_1 and W_2 representing the proton's structure become a function of the ratio ω concerning the square of the momentum transfer between the electron-proton (q^2) interaction and the electrons' energy loss (v). In conclusion, we can see how

Bjorken (Khalifa, 2017: 33) used the asymptotic explanatory pattern in his explanation of the scaling phenomenon (BSL):

Explanation Target:
Why does $W_1 = F_1(\omega)$ and $\nu W_2 = F_2(\omega)$ in diverse lepton-hadron scattering experiments?
Application of the Explanatory Pattern:
W_1 and W_2 can be expressed as mathematical functions of q^2 and ν.
E and θ are constant in the Bjorken limit in W_1 and W_2.
Given the underlying microphysics of the lepton-hadron scattering experiments, differences in E and θ are irrelevant to W_1's equalling $F_1(\omega)$ and νW_2's equalling $F_2(\omega)$ in the lepton-hadron scattering experiment.

In this representation of the Explanation Target E describes the beam energy of electrons and θ is the scattering angle of an outgoing electron. Due to the fact that W_1 and W_2 determine σ_{DIS}, there is a link between the scaling phenomenon and the unexpected scattering results. Before the scattering experiments it was already well known that:

$$\sigma_{DIS} \approx \sigma_{MOTT} \left[W_2 + 2W_1 tan2\left(\frac{\theta}{2}\right)\right].$$

Here θ is the scattering angle of an outgoing electron; and in fact, Bjorken's scaling laws implies that $W_2 + 2W_1 tan2\left(\frac{\theta}{2}\right)$ approaches 1 as the energy level of the scattering experiment is increased. Therefore, Bjorken's explanation of deep inelastic scattering results as consequence of the explanation of scaling phenomena.

According to Khalifa, to describe why Bjorken provides an advancement of our understanding of the scattering results, his *Nexus Principle* comes in help. Although he was right and thanks to the experiments of the SLAC-MIT team discovered the Bjorken scaling curves, researchers did not understand why SP and BSL were true. The analysis of Khalifa (Khalifa, 2017: 35) is based on the Outright Understanding (OU):

OU: "S understands why p" is true in context C if and only if S has minimal understanding and S approximates ideal understanding of why p closely in C.

According to him, Bjorken provided an explanation of the phenomena, but he missed to hit the contextually relevant benchmarks of the scientific community in which he worked. Khalifa suggests that the physicist community had two desiderata to be satisfied: (1) they wanted more information about the underlying microphysics of the scattering experiments, which is more explanatory information, as the NP requires. This was an important factor that Feynman indeed succeed in improving; (2) moreover, they

wanted to improve the scientific knowledge of the explanations provided, which is what the *Scientific Knowledge Principle* suggests. Physicists were not convinced that the experimental evidence was enough to accept Bjorken's asymptotic explanation. Still, according to Khalifa's NP, we can say that thanks to Bjorken's explanation our understanding of the phenomena in question is better than before, but he does not provide OU. In fact, physicists later on gained an even better understanding of the phenomena. Khalifa indeed proposes his EKS model to account for the most relevant features of the physicists' practices.

In this story, Bjorken plays an important role, in so far as thanks to his explanations we can have a basis to criticize some tenets of QED and so expanding the understanding of the scattering phenomena and subatomic particles. But according to Khalifa (2017) the explanations provided are incomplete and therefore they did not provide understanding to other physicists. Nevertheless, without his work, Feynman would have more difficulties in implementing Bjorken's explanation, in order to propose the mechanical model of partons, advanced in August 1968: its main idea is that hadrons are composed of hard, point-like entities, that Feynman called partons[32]. The parton model adds a mechanical interpretation that researchers found easier to understand than Bjorken's intricate mathematics.

[32] Now partons are called quarks. Back to the 60s the "quark" terminology was already whispered at the SLAC, but nobody would have talked out loud about quarks, as Arie Bodek admitted during his talk "Evidence for Fractional Charge of Partons" given in a Symposium organized by MIT on October 25 in 2019, to celebrate the 50 years of the quark discovery.

According to Khalifa, Bjorken's explanation are incomplete for two reasons: the first is that they fail to «provide values for all of the variables involved in an accepted or acceptable explanation schema, as is the case with Bjorken's failure to fill in the Asymptotic Explanation Schema describe earlier»; and the second is that «underlying microphysics is a variable in the Asymptotic Explanation Schema, but it would be hasty to infer from this single historical example that all understanding-conferring explanations must be microphysical, reductive, or mechanistic in character» (Khalifa, 2023: 36). These gaps in Bjorken's explanation were filled with the mechanical model proposed by Feynman, which used two strategies already in play when theorizing about the high-energy proton-proton interactions. As noted by Khalifa, the first strategy is labelled "working the infinite momentum frame": «Feynman imagined that, because of their high velocity, each proton would "see" the other as relativistically contracted along its direction of motion – roughly as a "pancake". Because the strong interactions between protons are of short range, each proton would also see the other as a frozen snapshot of its constituent particles» (Khalifa, 2023: 36-37). The second strategy is "impulse approximation", according to which, Feynman «assumed that within a single pancake (i.e. proton), partons did not interact with each other. Consequently, in strong interactions, each parton acts as an independent, quasi-free entity» (Khalifa, 2017: 37). Thanks to that model, Feynman completed the Bjorken's explanation in two ways (Khalifa 2017):

1) In deep inelastic scattering, the incoming electron emits a photon, which then interacts with a single free parton. Feynman assumes that partons are structureless and point-like. Consequently, the cross-section of protons in the deep inelastic region is similar to the cross-section of electrons in this region.
2) Feynman's explanation of scaling: W_1 and W_2 measure the distribution of the partons' momentum within the proton. In an interaction between an electron and a proton, the partons' momentum would be determined entirely by the momentum transfer between the electron and the proton (q) and the electron's energy loss (v). Consequently, W_1 and W_2 scale with functions of ω (E_2).

Feynman proposal of a mechanical model of parton takes this explanation target: Why does W1 = F1 (ω) and vW2 = F2 (ω) in an electron-proton interaction? Following the explanatory pattern we have seen above, the resolution would be (Khalifa, 2017: 38):

Application of the Explanatory Pattern: The electron-proton interaction is made up of electrons and partons. The electrons and partons are hard and point-like and collide with each other. Per the infinite momentum frame and impulse approximation, when hard and point-like objects collide, their momentum transfer is a function of ω.

W1 and W2 are momentum distributions of the partons in these collisions.

We can se now that both explanations, Bjorken and Feynman's are collectively correct, and this is a requirement of the EKS model. Moreover, since Feynman gave further explanatory information about the scattering results and the scaling laws, his contribution is interpreted by Khalifa as an application of the NP.

«The EKS Model also suggest that merely possessing a correct explanation frequently falls short of understanding – the explanation must be *known*, to wit in a way characteristic of scientific practice» (Khalifa, 2017: 40).
Feynman's amendments to Bjorken's explanation contributed to understanding by adding to our stockpile of correct explanatory information, but also by facilitating scientific explanatory evaluation.
The physic research at SLAC in the 60s and also at CERN in the 70s went on and physicists were able to implement the explanatory knowledge that EKS model identifies with knowledge.

5.7. Khalifa's Remarks to De Regt's Contextual Theory of Scientific Understanding

Together with CUP and CIT, the resulting account of SU proposed by De Regt is the Contextual Theory of SU (CTSU):

CTSU: A phenomenon *p* is understood scientifically if a theory *T* of *p* exists such that:

1) Scientists in some context C can recognize qualitatively characteristics consequences of *T* without performing exact calculations; and
2) The explanation of *p* by T meets accepted logical and empirical requirements.

According to Khalifa, CTSU captures well how Feynman's parton model gives us understanding of the phenomenon. Moreover, it is a rich account enough that it cannot be subsumed under the EKS model, nonetheless, according to Khalifa (2017) EKS model enjoys many advantages over de Regt's account, which are synthetically:

1) Better coverage of more instances of understanding than De Regt's;
2) EKS mode identifies the conditions under which qualitative reasoning in the absence of exact calculation is contributing to the value of understanding: the kind of qualitative

reasoning that is at the heart of CTSU is evaluated insofar as it allows to «SEE the nexus (i.e. to have scientific knowledge of an explanation). […] the value of qualitative insight (a shared language, visual images, etc.) is exhausted by its facilitation of good old-fashioned hypothesis testing» (Khalifa, 2017: 45);

3) CTSU has the problem of irrelevant insights: scientists may have the ability to recognize a new qualitative consequence *Q* of *T*, which has nothing to do with *p*, but fail to recognize that *T* explains *p*. According to CTSU this is sufficient to generate understanding of *p*. By contrast, Khalifa's NP overcomes the problem of irrelevant insights.
4) The problem of improbable explananda: explananda which laid epistemic and logical issues on De Regt's theory.

To detail a little bit deeply the point 4), let's focus on Khalifa's (2017: 46) suggestion to rewrite 1) in CTSU to overcome the problem of irrelevant insights in this way:

> 1*) Some scientist *S* (in some context *C*) can recognize *P* as a qualitatively characteristic of *T* without performing exact calculations.

This rewriting it is consistent with De Regt's description of Boltzman's SU of the properties of gases through kinetic theory (De Regt and Dieks, 2005: 152):

«If one adds heat to a gas in a container of constant volume, the average kinetic energy of the moving molecules – and thereby the temperature – will increase. The velocities of molecules therefore increase and they will hit the walls of the container more often and with grater force. The pressure of the gas will increase. In a similar manner, we can infer that, if temperature remains constant, a decrease of volume results in an increase of pressure. Together these conclusions lead to a qualitative expression of Boyle's ideal gas law».

What is here the subject of understanding are the relationships between temperature, pressure and volume that the ideal gas laws express. The SU is achieved by the inference of a qualitative version of the laws from a qualitative formulation of the kinetic theory, as it is the case following the 1^{*}). According to Khalifa (2017), the problem of improbable explananda regards two issues: one concerns the distortion of a semantic feature of the implications of CTSU, the other is related to the epistemic suspicion under which SU could fall. Regarding the first issue, if De Regt changed 1) with 1^{*}), he would be bound to accept this claim about explanation (ExC):

ExC: If *T* explains *P*, then *P* is a consequence of *T*.

This resulting claim about explanation is not free of well-known problems, which have been already criticized by Hempel. The famous example in that sense is that a person's having contracted syphilis explains why there is also a paresis issue. The problematic point is that only the 25% of people suffering from syphilis is also affected by paresis (Scriven, 1959). The problem is that it is possible to fail to recognize paresis as a consequence of a claim about syphilis. Issues of this sort are not addressed by De Regt (2017). He recognizes the similarities between his framework with the Hempelian deductive-nomological model, and albeit he and Dieks address the barometer and the flagpole problems (De Regt and Dieks, 2005: 162-3), they do not write about this issue. In response to the explanatory asymmetries we can find in the syphilis case, De Regt follows van Fraassen (1980) in stating that «it depends on the context whether the length of the flagpole makes it understandable how long the shadow is, or vice versa» (De Regt and Dieks, 2005: 164). If we appeal to the context in the syphilis case, it results that if *T* explains *P*, then *P* is a consequence of *T*; but this is not a matter of pragmatics, it is instead a matter of semantics. Appealing to the context is not enough to solve this problem. The consequence relation of the ExC does not require inference to context.

The second issue is connected to the former. In fact, De Regt could insist that the consequence relation is context-sensitive, but then understanding would become epistemically suspect: «If, in a particular context, one can "infer" that someone has paresis from the

fact that he has syphilis despite the low conditional probability, then recognizing a theory's consequences is little more than forming psychological associations with a theory» (Khalifa, 2017: 48). De Regt would have then the urgency to patch up this claim with his view that understanding is not a «psychological byproduct of scientific activity» (De Regt and Dieks, 2005: 138). This tension could me instead solved thanks to the Khalifa's NP principle: «A person's having syphilis *causes* him to have paresis, and this is widely regarded as the relevant notion of explanation in this example even if there is no further inferential relation» (Khalifa, 2017: 48). The explanatory nexus captured by Khalifa with NP entails a notion of explanation which does not raise the issues which affect De Regt's account.

According to Khalifa (2017), NP is an ameliorated principle than the condition 1) of CTSU and the SKP is accordingly an improvement over the condition 2) of CTSU. De Regt wants that the explanations achieving understanding satisfy the accepted logical and empirical requirements: they must satisfy consistency and empirical adequacy. According to De Regt (2017) the logico-empirical requirements are sufficient for SU; this means that as long as the accepted requirements are consistent with the best scientific methods in play, De Regt and Khalifa lay in good agreement. But there is yet a minor issue in which the two views do not fit. Khalifa claims that beliefs in understanding scenarios have to be safe, while De Regt seems to be not committed to this. Contrastively, Khalifa argues that

De Regt should be more committed to safety. He excludes the possibility that scientists could formed the wrong doxastic state about an explanation while satisfying the logical and empirical requirements De Regt cites. In a scenario like this, even if *T correctly* explains *P* in the actual world, *T* could nevertheless have *incorrectly* explained *P* on the basis of the same requirements. But a proposal like this is incoherent: «in science, the accepted logical end empirical considerations typically aren't ones that permit such scenarios because if an explanation could have easily been false given the logical and empirical criteria it satisfies, then one should not accept those criteria but should instead seek out more demanding criteria that will rule out false positives» (Khalifa, 2017: 49). If CTSU departs from the tenets of SKP, the logico-empirical constraints would result too modest to discriminate between correct and incorrect explanations.

To conclude, Khalifa's view about SU differs from De Regt's in that it owes the main framework to the received view, defining understanding as a species of knowledge, reducible to explanatory understanding. The NP and SKP are used to define SU to have a better coverage than CTSU and to avoid the problems of improbable and irrelevant explananda. In the next paragraph I will describe some issues related to Khalifa's account of scientific understanding.

5.8. Objections to Khalifa's EKS Model

The SU account proposed by Khalifa is a compelling alternative to De Regt's view for the scholars inclined to rely on ideas near to the received view that understanding is explanatory knowledge and to the ones more sympathetic to factivism or quasi-factivism about understanding. Even if it has many advantages than CTSU for this line of thinking, it is to be noticed that the strict constrain about the reduction of objectual understanding to explanatory understanding is problematic. Paradigmatically objectual understanding as it is called by many epistemologists when a subject understands a phenomenon (not just why *p*, but indeed *S* understands *p*), is covered by EKS model in Khalifa's framework. But as we will se later on in the next chapters, when scientists deal with scientific representation of a phenomenon from which they cannot gain so much explanatory information, although they gain representational understanding, which is the understanding of how a phenomenon is determined, organized, how distributed are its functions, and how are instantiated the internal relations of its features, the EKS model falls short of usability. So, EKS model is a relevant account for a definition of a general kind of SU, related to many scientific theories in many disciplines, when the scientific activities consist in explanatory information formation, extraction, evaluation and confrontation. But it does not work when the scientific activities concern a local kind of understanding, related to specific models, relevant and local

hypotheses, especially in scientific areas in which the methodology implied are of a mixed sort, i.e. the determination of molecular structures in bioinformatics.

Another interesting objection to EKS model comes from Hunt (2022; 2023) which distinguish, building on Khalifa's account, two alternative views: conceptualism vs explanationism. On the side of explanationism we find Khalifa, as a scholar of the received view, while Hunt argues for conceptualism. Explanationism (ExP) is so defined:

> *ExP*: all philosophical aspects of understanding-why are encompassed by an appropriately detailed account of the epistemology of scientific explanations.

So, views that want to contest the received view have to resolve the challenge of explanationism. One way to do that, according to Hunt, is to consider theoretically equivalent formulations: providing the same explanations, they can differ radically in the understanding scientists achieve through them. Hunt then uses cases of scientific practice in which two presentations of the same explanation lead to differences in understanding, i.e. Lagrangian and Hamiltonian mechanics as equivalent formulations that display two main sources of intellectual differences (Hunt, 2022: 5). In these cases, Hunt argues, equivalent formulations show that some differences in understanding-why do not reduce to explanatory differences. The

EKS model could encounter this problem also when addressing the explanatory integration issue, which will be discussed in Chapter 6.

To conclude, the main difficulties of EKS model are in the distinct character of understanding-how in representations and understanding-why in explanatory knowledge. In the next chapters 4 and 5 I will present two studies of representational models providing SU even if they do not hit the Khalifa's criteria for EKS.

Chapter 6:
Scientific Understanding and the Explanatory Integration in Cognitive Sciences

Scientific understanding in the field of cognitive sciences is a multifaceted concept that necessitates reflecting on the integration of various explanations. In this chapter[33], I argue that different kinds of explanations regarding cognitive sciences can be integrated into an account of explanatory scientific understanding, as proposed by Khalifa. Moreover, I propose that scientific understanding should be distinct from mere knowledge and should be conceptualized as a nexus of explanation. This chapter explores the theoretical foundations of scientific understanding, discusses different types of explanations in cognitive sciences, criticises a reduction problem in Khalifa's account and elucidates how these explanations can be

[33] This Chapter is the result of the presentation I gave during the workshop CIFMA 2023 at Eindhoven University the last November. I would like to thank all the contributors for the helpful discussion that let me ameliorate the arguments in Galli (2024a *forthcoming*).

effectively integrated to foster a holistic understanding of cognitive phenomena. Through an interdisciplinary approach, the aim of this pages is to enrich our evaluation of cognitive sciences and promote a more unified perspective on scientific understanding.

6.1. Introduction

Scientific understanding in the field of cognitive sciences is a multifaceted concept that necessitates reflecting on the integration of various explanations. In this chapter, I argue that distinct kinds of explanations regarding cognitive sciences can be integrated into an account of explanatory scientific understanding, as proposed by Khalifa. Moreover, I propose that scientific understanding should be distinct from mere knowledge and should be conceptualized as a nexus of explanation. In these pages I explore the theoretical foundations of scientific understanding, discusses different types of explanations in cognitive sciences, criticises a reduction problem in Khalifa's account and elucidates how these explanations can be effectively integrated to foster a holistic understanding of cognitive phenomena. Through an interdisciplinary approach, the aim of this chapter is to enrich our conception of cognitive sciences and promote a more unified perspective on scientific understanding.

Scientific understanding is a pivotal concept in the scientific domain and the specific area of cognitive sciences in the last decades

has tremendously increased our understanding of cognitive phenomena. Since cognitive scientists aim to give an account of our cognitive life, namely how our minds work, how is it possible to understand a language, and even how it is possible to understand why the sky is blue or why decarbonization is urgent to slow down the climate change, philosophers engaged in the study of understanding and scientific understanding should be interested in focusing on their theoretical frameworks and methodologies. This is indeed the case of Khalifa, Islam, Gamboa, Wilkenfled and Kostić (2022), which advance a way to illuminate the explanatory integration issue in the Cognitive sciences on the base of Khalifa's account of scientific understanding. Given the interdisciplinary origins of Cognitive sciences, we find between their theoretical posits and methodological tools. In particular, cognitive sciences to comprehend the intricate workings of the human mind must employ diverse explanations that span various disciplines, such as psychology, neuroscience, philosophy, and artificial intelligence. Tracking back to the development of the cognitive sciences reveals two main approaches towards the object of research and the methodologies implied: unification and pluralism. These two approaches reflect the philosophical view about the unity of science as debated in the Vienna Circle. According to Neurath (1937), a defender of pluralism, sciences should have been coordinated. Carnap instead argues that all sciences should be reduced to one grand unifying theory. According to Gentner (2019), at the

foundation of cognitive sciences, researchers were predominantly pluralistic, while views of reductive unity were prominent in the 1960s thanks to the manifesto of Oppenheim and Putnam (1958), but the received view on reduction cannot be applied to cognitive theories. In the last years, some scholars have argued that it is possible to attain a unified science of cognition «by showing how functional analyses of cognitive capacities can be and in some cases have been integrated with the multilevel mechanistic explanations of neural systems» (Piccinini and Craver, 2011). The crucial problem remains that we do not have an efficient account of what explanatory integration entails (Miłkowski, 2013, 2016). The canonical framework of cognitive science (singular) between the 70s and the 80s was given by computational functionalism. The first cognitive scientist to work on a solution to the issue of different explanations and the relations among them, he labelled it as the "level" of explanations, was Marr. His theory of vision is a paradigmatic case of cognitive science. He distinguishes three explanatory levels. At the higher level, it is defined as a high cognitive task as a function mapping some input over some output (this is the concept of computation defined as abstracting from the constituent operations). At the second level, the algorithm that computes the function is specified. The third level concerns what is called the neural hardware implementing the algorithm. This schema, albeit quite simple compared to the problem complexity, reveals the meaningful feature of classical cognitive science: its way of abstracting from the brain.

Marconi (2001) adds another level of explanation to Marr's vision theory: the program level. The notion of "level" will be later crucial; it is Craver (2008) who develops this notion and argues that the theories of psychobiological sciences entail structures and dynamics in which the levels of mechanisms are fundamental. The explanatory levels are related in a hierarchical order between mechanisms explained from the higher to the lower level, i.e. we have the level of the spatial navigation function, then the hippocampus mechanism generating spatial maps, neurons inducing long-term potentiation and finally, the activation of NMDA receptor (Craver, 2005). From one unique framework (the functionalist in Marr's theory) to the mechanistic framework of Craver's claim, the common trait is that a unificatory purpose holds in the first paradigm of cognitive science. A consequence of this way of dealing with cognition is that, even if they distinguish different levels of explanations and their relations, no different kinds of explanations are involved and consequently, there is no need for integration. But in recent years, it became clear that the plurality of sciences involved in the study of cognition is so broad and structured and the cognitive phenomena are so complex that the post-classical studies, through the vertical neuroscientific expansion towards the brain and the horizontal one towards the body and the environment, cannot be subsumed under the singular term of science, but it is a family of sciences which collaboratively work on cognition, whit different theoretical backgrounds and methodological assumptions and instruments. In this panorama, it is clear that

different sciences producing different kinds of explanations present the issue of how to integrate the explanations in order to assure a common scientific understanding (as a new basis for establishing a unified field of knowledge exchange, comparison and consensus) of the cognitive phenomena. Here, the scholars of understanding can come to help.

Khalifa's account of explanatory scientific understanding provides a framework to integrate these diverse explanations into a unified understanding. However, it is imperative to distinguish scientific understanding from mere knowledge and emphasize its role as a nexus of explanation. In this chapter, I delve into the theoretical underpinnings of scientific understanding, elucidate the diverse types of explanations in cognitive sciences, and argue for the integration of these explanations to facilitate an integrative perspective on cognitive phenomena.

6.2. Contrasting Khalifa's and De Regt's Accounts of Scientific Understanding

Scientific understanding (SU) is a multifaceted concept that has garnered significant attention among philosophers of science. Two prominent scholars, Khalifa and De Regt, have proposed distinct accounts of scientific understanding that illuminate different aspects of this complex phenomenon. In this discussion, we will explore and

contrast Khalifa's (2012, 2013a, 2013b, 2017, 2019, 2023) account of scientific understanding with De Regt's (2005, 2017) perspective, shedding light on their fundamental differences.

Khalifa's account of scientific understanding emphasizes the importance of integrating diverse explanations to achieve a holistic grasp of natural phenomena. According to Khalifa, scientific understanding goes beyond mere knowledge acquisition; it involves gaining insight into the causal-mechanical, structural, and functional aspects of a phenomenon. This perspective contends that understanding arises when we appreciate the interplay of these three explanatory components. Khalifa's framework is rooted in the idea that scientific understanding is not a mere collection of facts but a deeper comprehension of the underlying mechanisms, organization, and purpose behind natural phenomena.

Khalifa's account places a strong emphasis on interdisciplinary collaboration and the integration of different types of explanations from various scientific disciplines. It underscores the need to connect the dots between causal-mechanical, structural, and functional explanations, recognizing that a comprehensive understanding emerges when these facets are interwoven. It is one of the most demanding accounts of SU and it requires that experts evaluate their explanations according to the best available methods and evidence. There are three main principles supporting his model of SU, called Explanation, Knowledge and Science (EKS). Let's recall the main three principles he advances. The first is the *Explanatory Floor* (EF):

EF: Understanding why Y requires possession of a correct explanation of why Y.

Since scientific understanding can be subject to improvements, given its dynamical nature, Khalifa advances the Nexus Principle (NP):

NP: Understanding why Y improves in proportion to the amount of correct explanatory information about Y (=Y's explanatory nexus) in one's possession.

The third principle is the Scientific Knowledge Principle (SKP), which bounds the notion of scientific understanding to knowledge, namely scientific knowledge:

SKP: Understanding why Y improves as one's possession of explanatory information about Y bears greater resemblance to scientific knowledge of Y's explanatory nexus.

This last principle gives the idea that even the same explanatory information could be linked to different degrees of understanding, given the abilities and the information of relevant theories, models, empirical observation and experience scientists can have. Khalifa also gives a detailed definition of scientific knowledge of explanation (SKE):

SKE: An agent S has a scientific knowledge of why Y if and only if there is some X such that S's belief that X explains Y is the safe result of S's scientific explanatory evaluation (SEEing).

He concludes that thanks to SEEing and safety, the epistemological concept that requires an agent's belief to not easily have been given the way in which it was formed (Pritchard 2009), scientific knowledge of an explanation is achieved when «one's commitment to an explanation could not easily have been false given the way that one considered and compared that explanation to plausible alternative explanations of the same phenomenon» (Khalifa, Islam, Gamboa, Wilkenfled and Kostić 2022: 8). Khalifa and colleagues then argue that his framework of SU provides a «fruitful account how different explanations, such as the ones discussed above, can be integrated. The Nexus Principle is the key engine of integration» (Khalifa, Islam, Gamboa, Wilkenfled and Kostić 2022: 8).

In contrast, De Regt's contextualism cannot fit well to address the explanatory integration in Cognitive sciences. In fact, due to his Criterion of Understanding Phenomenon (CUP), supporting SU, he cannot account for the integration of different kinds of explanations. According to CUP: A phenomenon p is understood scientifically if and only if there is an explanation of p that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency. Given the multiplicity of the entangled kinds of explanations in cognitive sciences, we should

eschew De Regt's contextual theory, if we want to improve the explanatory integration in cognitive sciences.

De Regt's account of scientific understanding focuses also on the role of scientific models in the acquisition of understanding. He argues that understanding arises from an intimate familiarity with and a deep engagement in the use of scientific models. De Regt contends that scientific understanding is closely linked to our ability to manipulate, apply, and navigate these models effectively.

De Regt's account places less emphasis on the integration of different types of explanations and more on the centrality of models in scientific practice. For De Regt, understanding is intimately tied to our ability to make predictions, explain phenomena, and solve problems using these models. The more proficient one can employ a model to achieve these goals, the deeper one understands the relevant scientific domain.

The critical distinction between Khalifa's and De Regt's accounts lies in their conceptions of what constitutes scientific understanding. Khalifa's perspective emphasizes a broader understanding that integrates various types of explanations, fostering a comprehensive grasp of complex phenomena. In contrast, De Regt's account narrows the focus to the practical utility of scientific models in achieving understanding.

Additionally, Khalifa's account highlights the importance of interdisciplinary collaboration, encouraging the integration of insights from different scientific disciplines to enrich understanding.

While acknowledging the importance of models, De Regt's account does not explicitly emphasize interdisciplinary integration to the same extent.

In summary, Khalifa and De Regt offer distinct perspectives on scientific understanding, with Khalifa focusing on the integration of diverse explanations and De Regt highlighting the role of scientific models in achieving understanding. These accounts provide valuable insights into the multifaceted nature of scientific understanding and offer different lenses through which we can explore and appreciate the richness of this concept in the philosophy of science.

6.3. Kinds of Explanations in Cognitive Sciences

Cognitive sciences are a plural area of inquiry specialized in human and artificial cognition. To the extent of this chapter, the plural term "sciences" is chosen over the singular "cognitive science" to focus on the different scientific disciplines and methodologies involved in understanding the brain's functioning and the ability to produce the mind[34]. Since each scientific disciplines have its own methods and idiolects, they produce distinctive kinds of explanations, even if they aim to answer common questions about the brain and the mind, namely they all, as cognitive sciences, aspire to

[34] See Marconi (2001) and Legrenzi (2002) for a discussion about the differences between the use of the singular and plural terminology.

understand cognition. Each branch of cognitive sciences produces peculiar explanatory information, which shall be integrated to achieve scientific understanding of cognitive processes.

We can trace different kinds of explanations as an output of psychological, neuroscientific, linguistic, and computational research, each of them capturing information about dependency relations within the target system. Even if, since the early foundation of cognitive sciences, it was clear that the convergence of different research programs and their explanatory tool-box, had the same objective (Miller, 1979), not all the researchers are convinced about the integrability of different explanatory information. Those defending the integration of mechanistic explanations «will take integrations of these explanations to be able to explain cognitive competencies such as language production and comprehension, memory, perception, problem solving, categorisation, and reasoning; but also general, flexible behaviours and real-time performance, as well as the processes of learning and development that are characteristic of the human cognitive system» (Taylor 2019: 4575-76). One of the main issues concerning using different kinds of explanation in the multifaced area of cognitive sciences is that the struggle of finding a criterium to distinguish between what is an explanation and what is not is highly demanding. The possibility of having different explanations, explanantia, for the same phenomenon E, the explanandum, poses a problem for the unificationists who are in search of a unifying criterium for explanation. On one hand, the

unificationists wish to exclude some kinds of explanations from the role of explanans, so that they could find the explanation satisfying the criterium. On the other hand, pluralists allow to have many kinds of explanations, each one playing the role of explanans, offering different angles on the phenomenon E, which we can understand completely only if we integrate all the explanations at our disposal together. The latter rules out the unicity of the explanatory criterium the formers want to be satisfied by a single dominant explanation. Khalifa and colleagues (2022) explore the possibility of integration between mechanistic, computational, topological and dynamical explanations.

6.3.1. Mechanistic Explanations

Mechanistic explanations are the focus of the research of Bechtel and Richardson, 1993; Machamer et al., 2000; Craver, 2007; Illari and Williamson, 2010; Glennan, 2017; and Craver and Tabery, 2019. This kind of explanations is widespread in the cognitive sciences. There is no consensus on the proper characterization of mechanisms or how exactly they figure in mechanistic explanations, but one of the most used definitions comes from Glennan, which conceives mechanism as follows:

Glennan's definition of mechanism: A mechanism for a phenomenon consists of entities (or parts) whose activities and interactions are organized so as to be responsible for the phenomenon.

There are many cases in which this kind of mechanism is used in cognitive sciences, and one non-exhaustive example is the explanation we can found of action potential in neurons, also known as "nerve impulse" or "spike". Action potential occurs when there is a rapid rise and fall of the membrane potential of a certain cell. In particular, when the action potential travels down an axon, we can notice that the electric polarity across the membrane of the axon cells changes. In response to a signal from another neuron, sodium and potassium-gated ion channels open and close rapidly as the membrane hit its threshold potential. There is then depolarization and repolarization as the ion channels move into and out of the axon, creating a change in electric polarity between the outside and the inside of the cell. A mechanistic explanation of this phenomenon can be found as follows:

Action potential: a mechanistic explanation of this phenomenon specifies parts such as voltage-gated sodium and potassium channels. It describes how activities of the parts like influx and efflux of ions through the channels underlie the rapid changes in membrane potential. In this case, mechanistic explanations spell out the relevant

physical details. Hodgin and Huxley model is a major achievement that is not a mechanistic explanation of the action potential.

Marraffa and Paternoster (2013: 14) describe well how the account of explanatory integration given by Craver (2007) and colleagues such as Becthel (2009), entails a sort of "inter-level" mechanistic explanations. To spell out the notion of explanatory integration, Craver (2007) examines the development of the explanations of Long-Term Potentiation (LTP) and spatial memory. He distinguishes at least four levels. At the top of the hierarchy (the behavioural-organismic level) are memory and learning, which are investigated by behavioural tests. Below that level is the hippocampus and the computational processes it is supposed to perform to generate spatial maps. At a still lower level are the hippocampal synapses inducing LTP. And finally, at the lowest level, are the activities of the molecules of the hippocampal synapses underlying LTP (e.g., the N-methyl Daspartate receptor activating and inactivating). These are "mechanistic levels" or "levels of mechanisms": the N-methyl D-aspartate receptor is a component of the LTP mechanism, LTP is a component of the mechanism generating spatial maps, and the formation of spatial maps is a part of the spatial navigation mechanism.

Integrating these four mechanistic levels requires both a "looking up" integration, which will show that an item (LTP) is a part of a upper-level mechanism (a computational-hippocampal mechanism);

and a “looking down” integration, which will describe the lower-level mechanisms underlying the higher-level phenomenon (the molecular mechanisms of LTP). According to mechanists, this account is well-suited to define the explanatory integration in Cognitive sciences. Due to the redefinition of explanations under the label of information concerning mechanisms involved in cognitive phenomena, this kind of explanation gives rise to what Khalifa and colleagues (2022) calls the Mechanistic-Based Integration of explanations in Cognitive sciences.

6.3.2. Computational Explanations

Most prominent alternative to mechanistic explanations in the philosophical literature: they are considered a subset of functional explanations – explain phenomena by appealing to their function and the functional organization of their parts (Fodor, 1968; Cummins, 1975, 1983, 2000).

The functions to which they appeal involve information processing. In computational explanations, a phenomenon is explained in terms of a system performing a computation. A computation involves the processing of input information according to a series of specified operations that results in output information.

Many computational explanations describe the object of computation as having representational content, but some challenge this as a universal constraint on computational explanations.

6.3.3. Topological Explanations

In topological explanations, a phenomenon is explained by appeal to graph-theoretic properties. Scientists infer a network's structure from data and then apply various graph-theoretic algorithms to measure its topological properties, which are structural or mathematical properties of the system. In contrast to mechanistic explanations, they abstract away from particular details of causal interactions or mechanisms found in the phenomenon:

For instance, clustering coefficients measure degrees of interconnectedness among nodes in the same neighbourhood. Here, a node's neighbourhood is defined as the set of nodes to which it is directly connected. An individual node's local clustering coefficient is the proportion of edges within its neighbourhood divided by the number of edges that could possibly exist between the members of its neighbourhood. By contrast, a network's global clustering coefficient is the ratio of closed triplets to the total number of triplets in a graph. A triplet of nodes is any three nodes that are connected by at least two edges. An open triplet is connected by exactly two edges; a closed triplet, by three. Another topological property, average (or

"characteristic") path length, measures the mean number of edges needed to connect any two nodes in the network (Khalifa and colleagues 2022: 6).

This kind of explanation is used to picture how some system has the property to efficiently propagate information, as the case study of the nervous system of Caenorhabditi elegans show (Watts and Strogatz, 1998; Latora and Marchiori, 2001; Bullmore and Sporns, 2012).

6.3.4. Dynamical Explanations

When we have dynamical explanations, phenomena are accounted for using resources of dynamic system theory. A system is dynamical if its state space can be described using differential equations. The equations describe the evolution of the system over time. For example, in the case of dynamic explanations of bimanual coordination, the explanation rests on the fact that only the in – and – anti – phase oscillations of the index fingers are basins of attraction. Also this kind of explanation, while picturing some dependency relations between the features of the phenomena, is not of the mechanistic kind.

6.4. Scientific Understanding and Explanatory Integration

Scientific understanding is a multifaceted and dynamic concept that plays a central role in the field of cognitive science. The intricate workings of the human mind demand an array of explanations that span across disciplines, encompassing neuroscience, psychology, philosophy, and artificial intelligence. What emerges from the case studies in cognitive science is a compelling argument: scientific understanding can serve as a unifying framework that harmonizes the diverse kinds of explanations inherent in this interdisciplinary field. This notion of scientific understanding does not seek to reduce explanations to a singular, reductive framework but rather embraces the plurality of explanations, offering a holistic and integrative perspective. In examining the features of this integrative scientific understanding, we find that it is characterized by its depth, coherence, pragmatism, and its ability to promote interdisciplinary collaboration, ultimately enriching our comprehension of cognitive phenomena.

Khalifa and colleagues (2022) propose two main ways to integrate different kind of explanations in cognitive sciences: the Understanding-Based Integration (UBI) and the Mechanism-Based Integration (MBI). UBI is ultimately a new view about explanatory integration in Cognitive sciences, while MBI concerns the received view about explanatory integration aiming at unifying the different levels of explanation in a mechanistic one. According to Taylor

(2021) we should not accept to dismiss cross-explanatory integrations of mechanistic, dynamicist, psychological, computational and topological explanations in cognitive sciences, as instead some philosophers argue (Kaplan and Carver 2011; Miłkowski 2016; Piccinini and Craver 2011). Khalifa is also a defender of pluralism in cognitive sciences regarding explanatory integration and he argues for an account of explanatory integration based on SU. On the other hand, MBI provides that all models in the cognitive sciences are explanatory only insofar as they give information about mechanistic explanations. Against this, defenders of pluralism provide examples of putatively non-mechanistic explanations. In response, MBI philosophers use two strategies. The negative strategy consists in revealing that the putatively non-mechanistic explanation are no explanation at all (Kaplan 2011; Kaplan and Craver, 2011). The other strategy is assimilation and reveals the putatively non-mechanistic explanation to be a mechanistic explanation but with an elliptical nature (Piccinini 2006; Piccinini and Craver, 2011; Miłkowski 2013; Povich 2015; Hochstein 2016).

In recent years another area of inquiry has gained a central protagonism also in cognitive science, namely the study of the family of algorithms collected under the umbrella name “deep-learning”. These algorithms are run on machines to reach human-like abilities in many tasks. As we have seen in Chapter 5, deep learning models are emerging from the connectionist paradigm and are now basically

studied for engineering purposes, but they seem to be useful also for cognitive aims. According to Perconti and Plebe (2020), deep learning models pose questions that cognitive science should try to answer, such as why deep convolutional models that are disembodied, inactive, static and free of contextual awareness, seem to be the closest representation to the patterns of activation in the brain visual system (Perconti and Plebe, 2020). They argue that deep learning «can and should have its say in cognitive science» on the basis that «the engineering objectives of deep learning have been met with such success that, for the first time, we have artificial models performing complex cognitive tasks at human performance level. The era of toy worlds in which models are restricted to highly simplified versions of cognitive capabilities is over. We now have empirical examples of algorithms solving cognitive tasks at the full scale of complexity» (Perconti and Plebe, 2020: 2). The question is: can Khalifa's framework account for this turn?

The main flow I detect in Khalifa's framework to this extent is that the EKS model can account for the explanatory knowledge in the process of achieving understanding – and it also succeeds as an account of explanatory integration, when the items in question are explanations consistent with a broad theory of explanation Khalifa is entitled to. In the example suggested by Perconti and Plebe, we can figure out a case in which deep learning models of a phenomenon play the role of a representation of such phenomenon, and due to their features, researchers can achieve a representational

understanding of the target-system, i.e. the patterns of activation in the brain visual system. We could extract explanatory information from the deep learning models in question, but according to Khalifa's framework, the information will be not enough to account for explanations. Still, the models can play a relevant role in the process of understanding phenomena, even if the explanatory information researchers will gain from them is at minimum in comparison to the use of different experimental methodologies.

On one hand, we have Khalifa's account, which is broad enough to account also for the case of explanatory integration in the cognitive science. On the other hand, I still feel the pressure – if we want an account of scientific understanding sound enough to capture also what happens in the scientific research using deep learning models – not to limit the epistemic account of SU to explanatory knowledge. Relevant information is obtained from researchers through representations of the target system, developed through deep learning models. So, if we want to capture also this cases of SU in cognitive science, some other conditions must be put in play. Here I recognise the value of Khalifa and colleagues' argumentation concerning the analysis of explanations and the consistency of their view with explanatory integration in the cases they study, although I suggest that the following conditions have to be satisfied, in order to account for both explanatory and representational understanding: Understanding Pluralism, Coherence Across Explanations, Pragmatic Utility, Interdisciplinary Collaboration, and Non-Reductive Nature of

Understanding. I sketch these conditions in the following lines, suggesting that their satisfaction could improve Khalifa and colleagues' account of UBI, even if they leave some open problems.

6.4.1. Understanding Pluralism

As demonstrated in Cognitive sciences case studies, scientific understanding delves beyond surface-level knowledge. It encompasses the ability to penetrate the layers of causality, mechanisms, structures, and functions that underpin cognitive processes. For instance, when exploring the concept of mirror neurons, scientists go beyond the mere awareness of their existence and investigate the neural mechanisms (causal-mechanical explanations), how they relate to imitative behaviours (structural explanations), and why they evolved (functional explanations). One could argue that in each explanatory setting, a distinct species of understanding is achieved so that we can have, for example, a local understanding concerning the phenomenon related to the neural mechanism. Each area of cognitive sciences would then be entitled to achieve a distinct scientific understanding of the phenomena under scrutiny, given the specific explanatory information the scientific research produces locally. With "local understanding" I want to describe the case in which cognitive scientists gain a complete insight into the phenomena they study, providing a richer and more

nuanced perspective than mere factual knowledge. The locality here is designed by the specific features of the scientific research in the distinct area. In this way, we should end up not only with different kinds of explanations but with different kinds of local understanding specific to each area, and different understandings for different areas of cognitive sciences.

Can this pluralism of understanding be integrated in the same way we want to gain explanatory integration? As the plurality of explanations has been tackled by two mechanistic and understanding frameworks, can we account similarly for the plurality of understanding? Someone could argue that each area of inquiry in the cognitive sciences does not give rise to an instance of scientific understanding. This would obliterate the issue of accommodating different kinds of understanding. But if the EKS model captures the instance of understanding in a specific context, with defined explanations at disposal, I submit that in each area of inquiry, explanations come with a distinctive local understanding. This means that we also need a way to integrate the local understanding, not only with the explanations but with the general and broader scientific understanding, which is the result of the explanatory integration. It will be a pluralistic understanding in two ways: the first is that it must account for the local levels of understanding, and the second is that it concerns the plurality of explanations we want to integrate together. The relation between locals and general understanding

should be explored in more detail, as its interplay with integrated explanations.

To conclude, it is important to fill this gap in Khalifa and colleagues' account of explanatory integration in cognitive sciences to stabilise the definition of scientific understanding in this interdisciplinary research area.

6.4.2. Coherence Across Explanations

One distinguishing feature of integrative scientific understanding is its capacity to weave together disparate threads of explanation into a cohesive tapestry. Rather than isolating causal-mechanical, structural, and functional explanations, it seeks to align and integrate them. In doing so, it connects the dots and identifies the points of convergence and divergence within the explanations. This coherence fosters a more comprehensive and interconnected view of cognitive phenomena, highlighting the intricate relationships between different facets of understanding. Given the many kinds of explanations involved, the UBI proposed by Khalifa and colleagues should be able to satisfy the requirement of being coherent. It is a well-debated issue (Khalifa, 2016) in the epistemology of coherence and understanding. It is related to the coherence justification epistemology. On this theme, Elgin (2007, 34) writes: «[An individual] proposition derives its epistemological status from a suitably unified, integrated,

coherent body of information. This is the core conception of understanding […] And it is the conception of understanding that is closely connected to explanation». And also Kvanvig (2003, 192) says: «The central feature of understanding, it seems to me, is in the neighbourhood of what internalist coherence theories say about justification. Understanding requires the grasping of explanatory and other coherence-making relationships in a large and comprehensive body of information». They are the principal defenders of the strong view about coherence in understanding. Khalifa (2016) on the other side argues that the relation between understanding and coherence is shallow: «coherence is not part of the "core conception of understanding"». Similarly, while the "central feature of understanding" is in the neighbourhood of coherence, it isn't at home there. On my view, understanding is quasi-coherent: it walks like coherence and talks like coherence, but does not require a coherentist epistemology. According to Khalifa (2016), the improvements in understanding are not due to coherence, which is implied in the objectual understanding and not in the explanatory understanding. Still, the issue of a definition of coherence among the explanations, given their different journeys through the sciences remains one to be tackled.

6.4.3. Pragmatic Utility

Scientific understanding, in its integrative form, is pragmatically useful. The idea that understanding is a pragmatic notion is already embedded in De Regt's account, while Khalifa sides with the received view of understanding, conceiving it as bound to explanatory information. According to them, scientific understanding is not a purely theoretical construct but rather a skill or ability that aids researchers in making predictions, explaining observations, and solving complex problems. Considering the study of working memory and executive function, we can see that integrating insights from neuroscience, cognitive psychology, and artificial intelligence enables researchers to develop practical models that simulate and predict these cognitive processes. The scientific understanding scope is then not only determined by the phenomena under research, but its applications can range different scientific areas. Scientific understanding in this way is not only an ability, but it has also a pragmatic utility, which broaden the specific scope of understanding a specific phenomenon or problem. This pragmatic utility not only deepens our understanding but also allows for the application of cognitive science findings in practical domains like education, healthcare, and technology development. I think that Khalifa and colleagues' account of UBI should tell more about the pragmatic utility of local and general instances of scientific understanding.

6.4.4. Interdisciplinary Collaboration

Perhaps one of the most striking features of this form of scientific understanding is its ability to foster interdisciplinary collaboration. In the case studies mentioned earlier, the integration of neuroscientific, psychological, and computational explanations exemplifies how cognitive scientists from diverse backgrounds can come together to tackle complex problems. The exchange of insights and methodologies across disciplines enriches the overall understanding of cognitive phenomena. Moreover, it encourages researchers to embrace the diversity of explanations, recognizing that different disciplines bring unique perspectives and tools to the table. The program of establishing the UBI framework for cognitive sciences should also delve into the sociology of science, since the collaboration among researchers and scientific communities is at the core of the possibility of explanatory integration.

6.4.5. Non-Reductive Nature

Importantly, this kind of scientific understanding is not reductive. It does not seek to reduce complex cognitive phenomena to a singular, oversimplified explanation. Instead, it acknowledges the multiplicity of factors and dimensions that contribute to our comprehension of these phenomena. While it integrates diverse

explanations, it does so in a way that respects the complexity and richness of cognitive science, recognizing that no single explanatory approach can capture the entirety of the field. While scientific understanding comes with a context-sensitive nature, recognized also by Khalifa, knowledge as a de-contextualising device to structure information in a coherent, justified and approximately true form. Understanding is then a tool to get knowledge and to foster scientific knowledge in many areas of inquiry.

To conclude, scientific understanding in cognitive science provides a framework that unifies and integrates the diverse kinds of explanations inherent in this interdisciplinary field. It is marked by its depth, coherence, pragmatic utility, and capacity to promote interdisciplinary collaboration. This form of understanding does not seek to reduce cognitive science to a singular explanation but rather embraces the plurality of explanations, enriching our comprehension of the intricate workings of the human mind. It is a testament to the dynamic and evolving nature of scientific understanding, which continues to drive progress and innovation in the field of cognitive science. Given the fruitful connections made by Khalifa and colleagues, they should specify whether according to their view, UBI does or does not lead to new scientific knowledge. If the aim of explanatory integration is to ensure and expand the relevant scientific knowledge of cognitive phenomena, UBI should play an important role in affirming it.

6.5. Conclusion

The explanatory integration in cognitive sciences is a relevant issue, not only in the specific domain of cognitive researchers, but also for philosophers working on the notion of understanding and in particular scientific understanding. The UBI framework is promising and it depicts a possible development for the understanding studies, conceived more broadly as a philosophical and cognitive endeavour.

In conclusion, scientific understanding in cognitive sciences requires the integration of diverse explanations, as defended by Khalifa and colleagues' (2022). Khalifa's account of explanatory scientific understanding provides a valuable framework for achieving this integration. Still, their program lacks some clarifications about the relation between local and general understanding, the coherence of explanations coming from different scientific disciplines and the social interplay between research communities.

Moreover, it must be emphasized the distinction between understanding and knowledge. The notion of integrative explanations in the scientific understanding framework highlights the need for a broader description of cognitive phenomena. By embracing an interdisciplinary approach and displaying case studies, this chapter advocates for a more unified and structured perspective on scientific understanding in cognitive sciences, ultimately advancing our understanding of the human, animal and artificial mind.

Chapter 7:
Scientific Representation and Deep-Learning Models: the Case of AlphaFold

The scientific enterprise enriches the debate about models. In particular, in the field of structural biology, a new deep-learning neural network system called AlphaFold has been applied for many purposes. It allows us to predict a protein's structure with high accuracy. In this chapter, I will present the system in light of the discussion of structure representation and argue for a specific kind of representational relation holding between the predicted model structure and its target-system. By doing so, I will criticize the artefactual approach advanced by Knuuttila (2021) and present the features that characterize the predicted structures of AlphaFold as simulation models. In conclusion, I will argue for the representational accuracy of deep-learning models in order to hit scientific understanding of the phenomena under scrutiny.

7.1. Introduction

The notion of model is one with a wide polysemy within the sciences and philosophy. There is no unique conceptual framework and definition able to define all the models involved in scientific activities. There is no broad consensus on any unified account of models, as stated by Gelfert (2017), and it is considered an obvious consequence of this void to assume that «if all scientific models have something in common, this is not their nature but their function» (Contessa 2010: 194). Moreover, if this characterization of models as functional entities is accepted, we must then specify how the models work as «carriers of scientific knowledge» (Ducheyne 2008: 120).

One of the basic relationships between the model and its target-system (T) that has to hold, if the model must carry scientific knowledge, is the representation. My aim is not to advance a general theory of scientific representation, but to propose a definition of the representational relationship between the specific kind of models produced by the deep-learning neural network system AlphaFold (AF), and their T. In the first part of the chapter, I present the main positions and definitions of models as functional entities. This, however, is mostly a study about the semantics of the representational relationship between AF and its T, for it is on the basis of that relation that such models are carriers of knowledge. Examples of this relationship regard models of actual T, such as the double-helix model of DNA, or the Bohr model of the atom, i.e.

models that represent existing objects, and also models of potential (non-actual) T, as the examples of repressilators, synthetic oscillators and the ultra-Keynesian model analysed by Knuuttila (2021), i.e. models that represent objects not existing in nature. According to the representationalist view what we learn from models presupposes a representational relation, while according to the inferentialist view, the representational feature of models is decoupled from their capacity of carrying knowledge. I claim that the representational relation presupposes the epistemic function of models of both actual and potential T. Later on, I discuss Knuuttila's (2021) artefactual view of models. Moreover, I argue that the example of models of potential T does not invalidate the role of the representational relationship, and I discuss the contest of Critical Assessment of protein Structure Prediction (CASP) and AF. Furthermore, I submit that AlphaFold models can be interpreted as simulation models, even they do not carry explanatory information. In conclusion, I argue that they hold a kind of morphic representational relation with their T. The general aim of this chapter is to give one of the first contributions to expand a philosophical account of deep-learning models in general and AF models in particular.

7.2. A Taxonomy of Models

Models have a central role in sciences. Even if there is no consensus about their nature and qualifications, scholars have elaborated on three main areas: semantics, ontology, and epistemology of models. The first relates to what the models represent. The second concerns what the models are. The third focuses on the cognitive function modelers exploit for epistemological purposes. I will focus mainly on the first area, addressing namely the relation between the model and its target-system, specifically in the context of material, artefactual and simulation models, as they are tackled by Rosenblueth and Wiener (1945), Knuuttila (2021) and Durán (2018, 2020).

There are three main conceptions of the model–T relation: the similarity conception, i.e., models and their T are to some extent similar; the structuralist conception, i.e., models represent their T in virtue of a morphic relation between them; and the inferential conception, i.e., models as scientific representations have to be analysed in terms of the inferential function. Each conception offers different answers to certain problems. Moreover, we can distinguish the instantial view and the representational view. According to the former, models instantiate the axioms of a theory, that is composed of linguistic and mathematical statements. The representational view instead holds that it is rather the language that is connected with the model, while the model connects to the world «by way of similarity

between a model and designated parts of the world» (Giere 1999: 56). In turn, the representational view has an informational and a pragmatic version. The former conceives representation as «an objective relation between the model and its target, which imbues the former with information about the latter» (Gelfert 2017: 26). According to the latter, instead, it is not possible to «reduce the essentially intentional judgments of representation-users to facts about the source and target object or systems and their properties» (Suarez 2004: 768).

A further distinction can be drawn between substantive and deflationary accounts of representation. Substantive accounts aim for a robust explanation of the function of a representation in terms of a fundamental relation between a model and its target. Deflationary accounts, instead, settle for a light characterization of the functional unit of representational devices. We will see that while Knuuttila's proposal is pragmatic and deflationary, even though recognizes a representational function of models, the AF models are better interpreted by the representational, informational, and substantive view.

7.3. The Artefactual Account of Models

AF models, as representations of proteins, are a result of sophisticated techniques that make use of experimental data and

abstract models. The 3d structures of proteins predicted by AF recall the structure of material models of a DNA strand but with a digital suit. One of the first studies on the representational capacity of models has been made by Wiener and Rosenblueth (1945). They analyse the role of material models of phenomena in scientific research, stressing their advantage with respect to abstract models thanks to their representational features. They describe a material model as «the representation of a complex system by a system which is assumed simpler and which is also assumed to have some properties similar to those selected for study in the original complex system» (Rosenblueth and Wiener 1945: 317). The relation identified by the authors between the material model and the original complex system can be seen as a case of similarity conception. This view then contrasts Suarez's inferential conception. These models are intended to be approximations and "surrogates" (Rosenblueth and Wiener 1945: 320) for the real facts under observation. But models can represent also facts not already present in reality. Indeed, Knuuttila is interested in developing an account of models consistent with the need, in some areas of inquiry as economics or synthetic biology, to build models of objects we do not find in nature or in society, i.e. models of invented objects.

Knuuttila (2021) advances the artefactual account of models which fits well with the inferential account developed by Suárez (2004). She is interested in stating an alternative position to the received ones, both substantive and deflationary, pointing out that

models can be carriers of scientific knowledge even if they do not represent the actual state of affairs in the world. She insists on the modal reach (Godfrey-Smith 2006) and the modal dimension of modelling (Le Bihan 2016), «which approaches models as purposefully constructed systems of interdependencies designed to answer some pending scientific questions» (Knuuttila 2021: 5). Models as epistemic artifacts function as "erotetic devices" (Knuuttila 2021: 6). Such devices are artificial systems that deploy dependencies constrained to the aim of answering a specific scientific question, supported by theoretical, and empirical considerations.

Two examples are described, one of an ultra-Keynesian model as an example of an economic model that does not refer to a real T, and one of repressilators and synthetic oscillators in synthetic biology, that do not correspond to any existing circuits, but are rather pictured to explore and test possible biological circuit designs. To strengthen the cases, she distinguishes between representational modes and media, and also between internal and external representations. The representational modes are the many semiotic devices that express various meanings and contents, while the representational media are for example the ink on paper, digital computer, biological substrata, and what support the representations. According to Knuuttila (2021: 5) the same representational mode can be implemented in different media as the example of the synthetic repressilator and the electronic repressilator that instantiate both the same ring oscillator design, yet

they are implemented in different media «enabling different kinds of inferences» (Knuuttila 2021: 5). Moreover, an internal representation concerns «how various kinds of sign-vehicles or representational devices are used to make meaning and convey content» (Knuuttila 2021: 5), i.e. for a material model of the atom, the material, the proportion, and in general the semiotic and semantic features of the model chosen to represent the specific object; by external representation, instead, she refers «to the relationship of a model to a real-world target system, the question on which the philosophical discussion has largely concentrated» (Knuuttila 2021: 6). This distinction is particularly relevant for the definition of models as epistemic artifacts: «Nevertheless, the fact that something may be internally represented within a model without necessarily representing the actual state of worldly affairs opens up the prospect of conceiving modelling as a practice of exploring the possible» (Knuuttila 2021: 7).

The artefactual approach allows us to see biology as a discipline that not only focuses on natural organisms but includes also potential organisms (Elowitz and Lim, 2010, 889). So conceived, models are carriers of knowledge in virtue of their being erotetic devices and artefactual constructs useful to support surrogative inferences about a potential target-system. In such a way, inferentialists would argue that their representational capacity is not relevant to their use in exploring the possible.

7.4. Some Remarks on the Artefactual Account of Models

The artefactual account stresses the pragmatic goal that directs the models' construction and manipulation. It is to conceive models as tools for investigating specific phenomena, used to answer scientific questions, motivated by theoretical, and empirical tenets. According to Knuuttila (2021), their accomplishment relies on their modal function of exploring the spaces of possibilities and the main point is that their success needs not be grounded on the representational relation between the model and the target-system. Thanks to the distinction between internal and external representations, Knuuttila safeguards a slightly deflationary definition of representation, which connects the artefactual models with a possible organism. Obviously, the correctness of models of merely possible T does not need the same kind of warrants as the models of real T. What does then warrant them? For Knuuttila it is simply their predictive success, without any need to invoke to any representational relation, yet it remains unanswered the question concerning what warrants the models' success. In other words, how can we probe the success of a model of a potential target-system, without any reference to the representational relation between the model and the possible state of affairs? Knuuttila (2021) claims that it is still sufficient for a modal relation to justify the success of the artefactual models[35].

[35] Knuuttila's approach (2021) and case study are similar to Cornelissen and De Regt (2022) about synthetic chemistry; see also Broeks, Knuuttila and De Regt, forthcoming. According to Cornelissen and De Regt (2022) we can dismiss the

We can reframe the modal feature of the relation between the models and the potential T as a predictive relation, i.e., a model would predict the possible state of affairs, if there were conditions such and such. One of the kinds of models so far used to explore the possible phenomena within a manifold scenario is the simulation model (SM). That is a model resulting from computational procedures able to predict or determine specific output with a given set of data. SM are helpful to study and predict complex scenarios and phenomena. They are implemented by a certain degree of idealization and can be used to study actual T (like biological systems, i.e. birds flocks, ant colonies, structure determination, enzyme kinetics and molecular dynamics) and potential T (like the behaviour of mechanics and artifacts as airplanes, spacecrafts, biomedical robots, and also new proteins, new drugs and possible organisms). As it happens with imaginary economics, repressilators and oscillators, from a set of data and techniques the respective models predict how the possible systems would act. To this extent, artefactual models are a kind of simulation model: though the examples are not strictly speaking computer-based simulations, they

---

factivist criteria about theories and models to be used to achieve SU; SU entails a non-factivist account of theories and models. Indeed, they defend the idea that SU requires intelligibility, namely that theories and models used have to be intelligible, rather than representational accuracy. While I argue for a factivist account of representation in a realist way, Cornelissen and De Regt propose an antirealist account of intelligibility, coherent with De Regt's theory of contextual scientific understanding.

simulate possible states of affairs, useful to predict how the system will work.

I submit, however, that neither for simulation models nor for material models we can easily dismiss the representational link between the model and its T. In the case of artefactual models, it seems intuitive not to stress the representational link, because we weigh differently the conceptual role of an actual T and a potential T. However, if we want to gain epistemic access to the T in question, actual or potential, the model has to maintain a representational link with it. I call it the accessibility condition (AC):

Accessibility condition: A model M of a target-system T is a functional carrier of knowledge in virtue of its capacity to give epistemic access to T through the representational relation established by the researchers between M and T.

In the case of AF the output models of predicted proteins' structures can be conceived as a kind of artefactual model. Most AF models represent actual target-systems, but they are also useful in the exploration of potential proteins. In that case, their success depends on their accurately representing the modal properties of proteins, i.e., what is actually possible or impossible for proteins. The discussion on representation, then, is far from over, and a substantive view of representation is still in play.

7.5. CASP and AlphaFold Protein Structure Prediction

AF is a breakthrough deep-learning network AI system able to predict highly accurate protein structures. Its computational power and sophisticated engineering let the DeepMind team, which worked on it, win the CASP 14 (Critical Assessment of Protein Structure Prediction) on the 30th of November 2020. CASP started in 1994 and it is a biennial competitive appointment for biological researchers working on protein structure prediction, aiming at solving the well-known folding problem: How is it possible to fold a protein starting from its strains of amino acids? The founder and chair of CASP is John Moult, Professor of the Institute for Bioscience and Biotechnology Research and the Department of Cell Biology and Molecular Genetics at the University of Maryland. He describes CASP in this way:

Computational biology differs from traditional science in that it takes place in a virtual world. Achieving rigor in a computational world which the scientist controls is much harder than when dealing with the inflexible realities of the physical world. We introduced Community assessment experiments in computational biology to help achieve the same rigor as in real world science. CASP (Critical Assessment of Structure Prediction), the first framework for these experiments, is an organization that conducts double blind community wide experiments to determine the state of the art of computational methods for modeling protein structure from amino acid sequence and other information. CASP has now been running for over 20

years, with continuing high participation rates (over 100 groups around the world), and has been accompanied by an enormous improvement in the accuracy of the protein modeling methods. The CASP methodology has now been adopted in a wide range of computational biology areas, including protein-protein interactions, genome sequence annotation, biological networks, and protein function annotation (Moult 2022).

The first lines make a sharp distinction between the rigor achieved in the real-world sciences and the one obtained in a computational world. I am interested in showing the philosophical relevance of the effort to make the two methodologies meet and enhance each other. Two questions. Why do the real-world sciences working on protein folding need such an upgrade? Moreover, why is it so important to solve the folding problem? “We have discovered more about the world than any other civilization before us. But we have been stuck on this one problem. How the proteins fold up. How the protein goes from a string of amino acids to a compact shape that acts as a machine and drives life?”, says John Moult (2021), filmed in AlphaFold: The making of a scientific breakthrough, the inside story of DeepMind research team who created AF. This is indeed the folding problem. Solving it means making huge steps in molecular biology and consequently in many other biological fields. DeepMind team states that the research program that leads to AF and similar systems is crucial for the development of the life sciences. Proteins are stunning biological nano-machines, whose understanding will take us to unveil how they work and interact with other molecules.

They are polymers in which the 20 natural amino acids are connected by amino bonds. They are polymers in which the 20 natural amino acids are connected by amino bonds. They are synthesized by the ribosomes, which are complex molecular machines present in all living cells, measuring around 30 nm. Ribosomes compose amino acids together in the specific order defined by messenger RNA molecules.

AF team trained this system on publicly available data consisting of around 170.000 protein structures taken from the protein data bank (PDB), together with large databases containing protein sequences of unknown structure. Thanks to the genomics revolution we can read amino acid sequences of proteins at massive scale; in fact, the Universal Protein database (UniProt) contains 180 million protein sequences. The building blocks of proteins are amino acids, small molecular compounds with unique features composed of between 10 and 20 atoms. In ordinary biology we find 20 standard types of amino acids floating within the cytoplasm of the cells. They connect to a piece of transfer RNA that matches with the three genetic sequences of the genetic code of the RNA messenger. Ribosomes then read the three-basis instructions of the RNA messenger and start building a chain of amino acids that goes out from the ribosome. As the chain of amino acids exits the ribosome, released in the cytoplasm, it is surrounded by water molecules and subject to the interaction of physical forces that make the chain fold up on itself and form the complex 3d structure we call a protein. All this process

is called translation because the molecular mechanisms manage to produce a fully operative protein with proper functions by translating a piece of the genetic code. The unique shape of a protein is defined by its amino acid sequence and its shape is the key to unlock its functions. Determining the 3d structure of a protein is indeed necessary to understand its functions. Proteins seem like pieces of a puzzle, but with a dynamic shape which can change according to the bonds they make with other interacting molecules. Nonetheless, a protein would bond with some molecules and not others. There are specific combinations of proteins and molecules. By understanding the protein shape and the occurring molecular interactions, scientists can design vaccines, new drugs and functional structures for ecological purposes: "Among the undetermined proteins may be some with new and exciting functions and—just as a telescope helps us see deeper into the unknown universe—techniques like AlphaFold may help us find them" (The AlphaFold Team 2020).

Proteins are fundamental for most living beings, and enhancing their understanding through computational allows us to tackle diseases, discover new medicines and disclose the enigmas of life in a faster and cheaper way than traditional research on existing proteins. Thanks to painstaking experimental effort, real-world sciences have determined before the release of AF the 3d structures of approximately 100.000 unique proteins (Thompson, Yeates and Rodriguez 2020; Bai, McMullan and Scheres 2015; Jaskolski, Dauter and Wlodawer 2014; Wüthrich 2001). Using the experimental

methodology scientists had at their disposal until now, it could take from months to years and a lot of financial resources to determine a single protein structure. Computational methodologies are in fact needed to reduce this gap and to "enable large-scale structural bioinformatics" (Jumper, Evans, Pritzel et al. 2021a: 1). That is why CASP has been promoted within the biological fields, with the aim to push researcher communities to solve the protein folding problem, that has been an open research problem since when, around 1960, the first atomic-resolution protein structures were proposed (Kendrew 1961; Pauling and Corey 1951; Pauling, Corey and Branson 1951), while the first protein structures detected presented unpredicted irregularities. It was the case of globin structures, a clade of globular proteins containing heme, a precursor to hemoglobin (6,5 nm), involved in binding, and transporting oxygen. Globin proteins contain the globin fold, which is a series of eight α-helices packed together in irregular ways. Since the 60's the folding problem concerns three different problems (Dill, Ozkan, Shell and Weikl 2008):

1) The folding code: the thermodynamic question of what balance of interatomic forces dictates the structure of the protein, for a given amino acid sequence;
2) Protein structure prediction: the computational problem of how to predict a protein's native structure from its amino acid sequence;

3) The folding process: the kinetics question of what routes or pathways some proteins use to fold so quickly. We focus here only on soluble proteins and not on fibrous or membrane proteins.

The main CASP evaluation follows the criteria of comparison between the predicted model α-carbon positions and those in the real-world target structure. The visualisation of cumulative plots of distances between pairs of α-carbon in the model and target structure positioning is used to evaluate the prediction against the experimental result, such as shown in the two figures aligning computational prediction with the experimental result. The real structure is already known by the evaluator so that the CASP examination can estimate the accuracy of the predictive model. To each prediction is assigned a numerical score GDT-TS (Global Distance Test—Total Score) specifying the percentage of modeling residues in the model with respect to the target.

The CASP campaign evaluation relies basically on the issues of 1) The folding code, 2) Protein structure prediction, and 3) The folding process, although the results are carried out in many prediction categories: tertiary structure prediction, residue-residue contact prediction, disordered regions prediction, function prediction, model quality assessment, model refinement, and high-accuracy template-based prediction. Tertiary structure prediction is then divided into three sub-categories: homology modeling; fold recognition; and de novo structure prediction (New Fold). All these conditions form what

we can call the accuracy qualification (AQ). The higher the GDT scores, the better the AQ of the predictions, and the higher the AQ, the nearer the model to the real shape of the protein. Another consequence of the AQ is that higher scores correspond to higher amounts of correct information transmitted from T to M, and from M to the modelers.

Since 2018 CASP team made some improvements, but the big leap was between AlphaFold 1 (AF1), the ancestor, and its successor, AlphaFold 2 (AF2), whose score, according to Moult, was around 90 GDT on 100 points scale prediction accuracy. DeepMind developed new deep learning (DL) architectures to improve the research methods for CASP14, which led to a high level of accuracy. These methods are inspired by the research areas of biology, physics, and machine learning (ML) and by the studies many scientists enhanced during the years on the protein folding problem. The AF2 system is described as a neural network-based model (Jumper, Evans, Pritzel et al. 2021a). It is important to note that it is described as an AI system coherent with the wider project of Demis Hassabis, CEO and co-founder of DeepMind, of making further steps in General AI. The whole AF architecture learns from the data and elaborates the 3d structure prediction of the folded protein. We can think of a folded protein as a spatial graph, a spatial presentation of a graph in the 3-dimensional Euclidean space R3, in which residues are the nodes and edges link the closely related residues (Jumper, Evans, Pritzel et al. 2021a). The graph matters to understand the proteins physical

interactions and their evolution. For the second version of AF2, the team created an attention-based neural network system, trained end-to-end, that attempts to interpret the structure of this graph while reasoning over the implicit graph that it's building (Jumper, Evans, Pritzel et al. 2021a). By process iteration, AF2 produces accurate predictions of the underlying physical structure of the protein in days-time. Moreover, the system can predict the reliability of parts of each predicted protein structure using an internal confidence measure. The following is the AF1 architecture that provided important results in CASP13, beating the median free-modelling accuracy of other systems.

AF1 has a straightforward architecture (Senior, Evans, Jumper et al. 2020). It begins with the amino acids sequence for which we are searching the protein structure. The first step concerns a data extraction move from the known database, in order to find similar protein sequences. The first task of the neural network is to find similar sequences, and it is called Multiple Sequence Alignment (MSA). The protein structure is responsible for its function, and we know that evolution carved the organisms in such a way that only some structures passed the survival threshold. Indeed, in different organisms during evolution a protein structure is more stable over time than the genetic sequence encoding that particular protein the genetic mutations that passed the evolutionary test are those that did not affect the protein structures. Comparing evolutionary-related protein sequences, whose 3d form should share some similarities, is

what MSA does: scrolling the database to find amino acid sequence matches in the animal kingdom. To sum up, in AF1, 3 main steps need to aim at structure prediction:

1) AF1 collects the MSA features;
2) it predicts then the distogram using a residual neural-network;
3) it optimizes the protein backbone using the predicted distogram in combination with simulated physical forces. The output is the 3d predicted protein structure.

As the aforementioned system, AF2 presents three main blocks:

1) A pre-processing stage where the input sequence is used to query additional information about the initial sequence from databases;
2) The information is then mapped into an MSA and pair representation, which are refined by the Evoformer, a 48-layer deep transformer-like network that uses attention mechanisms to update MSA and pair representations;
3) The structure module, a recurrent network, processes the Evoformer output, which transforms the abstract representations of the Evoformer into concrete 3d coordinates of the protein geometry.
Just to cite the improvement of the new architecture of AF2, it allows for jointly embedding of multiple sequence alignments (MSAs) and pairwise features. Moreover, AF2 has a new representation output

and an associated loss that together allow for end-to-structure prediction.

The AF research teams does not submit that the system is capable of revealing underlying laws regulating protein folding. AF, however, seems to have reached important results concerning some kinds of proteins, especially those based on a strain of between 100 and 200 amino acids. Moreover, albeit the neural networks system distances the empirical link of evidence gathered from experimental data in the genomic database of proteins, it has the computational power to disclose the structure of the simulated object. In future, it may be capable of finding common patterns between the structures predicted. In any case, from a philosophical perspective, it is important to ask whether this kind of AI system can assist researchers in unveiling recurrent structures that could be defined as the laws governing protein folding. This discovery could improve even better the system solving the folding problem.

7.6. AlphaFold as a Simulation Model

In the last years, as the use of deep-learning neural networks has become pervasive in engineering and scientific areas, scholars have focused correspondingly on the diffusion of simulation models as tools and outputs of neural network systems. What are simulation

models is then a crucial issue in the epistemology of models and the general philosophy of sciences.

A simulation model (SM) is a representation of a real or possible system, interacting with a determined environment, supported by computation techniques and expressed through visualization tools. It is a powerful instrument to represent, observe, study and manipulate to a higher degree of realism complex phenomena within a system. I submit that a model produced by AF is a kind of SM endowed with a degree of accuracy that was not available in the past, therefore improving the representational link between M and the related T. I submit that AF is a system architecture that produces SM of proteins' structures. We can divide AlphaFold into three main sectors: 1) AF as a complex neural-network system as a whole architecture; 2) AF sector sequences of algorithmic processing, the main blocks of the architecture; 3) AF's protein structure model as the output of the system.

As we know, the first stages of the system have to do with the analysis of the protein structure data contained in the database. In fact, in CASP the accuracy of the predicted structure is measured through the structure model obtained via experimental methods through X-ray crystallography and NMR spectroscopy. I claim that in each sector AF works as a kind of SM. According to CASP14 there are three relations to be noted:

1) The first between the real target system T and the experimental model, i.e. the relation between the real receptor-binding protein adhesin (Fig. 1) and the model resulting from the use of X-ray crystallography and NMR spectroscopy;
2) The second between the experimental model and the simulation model AF, namely what is pictured in Fig1, the relation between the model obtained through experimental methods and the simulation model produced by AF;
3) The third between AF, the whole system architecture and the real target-system T, i.e. the real adhesin protein. The experimental and simulation success of these models is due to the relation they have with T.

In the first case, the relation is obtained through experimental work which preserves the empirical link between observation and data manipulation. In the second case, the two different kinds of models are both successful representations of T, even though the simulation success entails a higher abstraction than the empirical link the experimental model holds in the first place. In the third case, AF as a whole architecture and the target system are not linked by an empirical relation, in so far as there is no direct observational contact as in the case of X-ray crystallography or NMR spectroscopy between the enquirer and the T. They are connected through the data manipulation and the simulation process binding the initial data, with the structure model in output.

Given the digital, computational, and algorithmic nature of the AF system, we can interpret it as an architecture producing simulation model (SM). There are mainly two types of simulation models: 1) SM is conceived as an implementation of models already existing; for example, aerospace engineers use SM of planes to test models they already have under specific circumstances like mechanical stress and weather conditions; 2) CS as models which have their own complexity and autonomy, the study of which is enhanced focusing on computer science and software engineering. According to Durán (2021: 317), a simulation model (SM) is a "rich and complex structure that departs in important ways from standard models used in scientific research". Furthermore, Durán (2021) argues that the construction of the SM is possible because of a new methodology that is in place. He calls it recasting, and it consists of clustering a multiplicity of models into one fully computational SM. Think of it as the mashing-up of different models, also theoretical and mathematical, that could be implemented through deep-learning networks, with the specific aim to predict, in this case, the folding of proteins. To refine the terminology for the purposes of AF, we can call the methodology in place reshaping. AF begins with a set of data with empirical and experimental information, then through the intervention of programmers in adjusting the learning bias with respect to the desired output, using different integration modules, idealisations, and reshaping the data representation with the multiple sequence alignments MSA, according to cycles of implementation

and integration, through the Evoformer and the Structure Module, we gain the visualization of the 3d geometry of the folding shape of the protein.

Not all the SM produced by AF are accurate representations of their T, especially the complex proteins are very hard to predict through the AF architecture as it is. Moreover, AF does not predict important aspects of protein structures as many ligands, metal ions and cofactors. Furthermore, the main limitation of AF is that the system predicts only a single state of the protein, and it is also hard to tell which state of the protein will be represented by the model (Perrakis and Sixma 2021). In fact, AF produces indeed SM with specific aims and empirical and theoretical assumptions and limitations, that must pass the abovementioned accessibility condition AC. Moreover, given the accuracy standard gained from the experimental data, we can draw another requirement to be satisfied, the correctness condition (CC) for the proteins models:
Correctness condition: SM represents correctly iff the accuracy qualification (AQ) is satisfied.

The AQ developed by CASP is a threshold for the correctness of the representation of SM. I take it as the level of approximation to reality the representation gains from the system through the work of modelers.

To conclude, AF consists not only of a complex and sophisticated computational implementation of the experimental models of

proteins' structure determination, but it is a simulation model which is already changing the scenario of the computational and structure biology research areas.

7.7. Structure and Representation

I have advanced an interpretation of AF models as simulations. Thanks to the simulation power, modelers have greatly improved the representational capacity of models. Now I suggest a definition of the relation, refined through simulation, holding between the AF models and the objects they aim to represent:

Structural Dynamic Approximate Isomorphism: a mapping that gathers through simulation even more information about the dynamic structure of T, so that the two systems (the model and T) approximately share the main structural features.

This definition pictures the ideal isomorphism between the model and the real protein which AF assumes as an implicit presupposition. It is a form of mapping since AF aims to visualize the shape of the protein as an image which can be navigated and observed in many aspects on a computer. The two systems should share the same features, represented one-to-one in the model: the individual folding units (domains), dynamic movements, contact matrix, ligands, and

each polypeptide chain, and monomers, involved in multimers. Moreover, the two systems should share the same features under the same dynamics, i.e. the interactions of the domains in T should correspond in the mapping of the model. Given the limitation of AF, the definition assumes that the simulation model could be refined through time thanks to more and better information about the relevant features of the real proteins. The isomorphism between the two systems should regard the geometry as the information detected regarding the ligands and the folding units. In the case of protein folding the isomorphic relation is fundamental between the two systems, in so far as the protein shape is responsible for its function.

Why should the isomorphism be dynamic? One of the most important limitations of AF is that it predicts only a single state of a protein, but the aim of the AF researcher is to overcome this boundary. AF models are the peak of an important history of views about, and scientific representation of, proteins. In the last century structure biologists shifted from the static view, according to which the protein models represented rigid structures, to the dynamic view:

> The study of how proteins serve the needs of a living organism is a curious case in which a method that yielded dramatic advances also led to a misconception. The method is X-ray crystallography […] The intrinsic beauty and the remarkable detail of the structures obtained from X-ray crystallography resulted in the view that proteins are rigid. This created the misconception, namely that the atoms in a protein are fixed in position (Karplus and McCammon 1986: 42).

The dynamic turn in protein representations owes a lot to thermodynamics. In fact, the dynamic analysis treats proteins as thermodynamic systems. The shift brought changes also to the structural concept. The old structural concept, coherent with the static view, is committed to the beliefs 1) that every protein has a rigid and static 3d structure and 2) that the protein structure alone determines protein function. The new dynamic concept of protein structure drops these commitments and adopts an inferential stance toward the proteins' structures, which are taken to be flexible, dynamic and constantly under structural fluctuations and mutations according to the environment and occurrent phenomena. Advocates of the dynamic concept are committed to the belief that dynamics and structures are relevant determinants of protein behaviour and function (Neal 2021). The supporters of the dynamic concept suggest a wide range of experimental, theoretical and computational strategies to test the dynamic properties of proteins. AlphaFold researchers support the dynamic view of protein structure, well represented by accurate prediction models.

The motivation of AF is that biological research will be aided by the availability of an open-source determination structure database. The assumption underlying AF system and fostering this motivation is that simulation model structures entail an isomorphic relation with the target-protein. The protein may be in the real world, or a possible protein, or a protein mutation, whose structure is to be explored, in order to accomplish some specific functions, as in the case of PET

depolymerization (Lu, Diaz, Czarnecki et al. 2022). AF model assumes that the dynamic view can be fostered through computational methods via deep-learning network architecture.

The AF system architecture is built to replicate the shape of the proteins according to their geometric features. The SM is apt to replace the representation of a protein given by the experimental procedures. The accuracy of the AF models is then grounded on the approximation to the structure of the real protein or to the functional structure of potential proteins. What best captures the conservation of information and geometric features between M and T is the notion of isomorphism. Related to protein structure prediction or drug discovery, AF researchers are therefore committed to a kind of isomorphism. On its basis, we can then define the representation relation:

Representation: A scientific model M represents a T, which may be actual or potential, iff the dynamic structure of the model is approximately isomorphic to the structure of the T.

This kind of definition avoids some problems described in the structuralist conception of scientific representations[36]. According to Suárez (2003) and Downes (2009) isomorphism cannot ground the

[36] This view is consistent with the factivist stance defended by (Rice, 2016), according to which idealized models (and such AF models are algorithmically idealized) can provide factive scientific understanding, even if they do not purport an accurate scientific representation of the difference-making features of the target-system in the world.

representation relation, because the former is characterized as reflexive and symmetrical, while the latter is not. Frigg and Nguyen (2017: 55) coin the requirement of directionality to account for this asymmetry. To answer these critics, let us recall that AF modelers do not aim at ideal models of proteins. The 100% GDT score is an ideal limit of research output, while the condition to be obtained is the standard of accuracy, i.e. AF models are accurate in so far as they represent their T, as an experimental representation of them would have done. The accuracy of AF models relies on the training the networks have got from the experimental data gathered. The isomorphic relation is approximate in the sense that the relation safeguards the correctness condition (CC).

Moreover, since the function of a protein depends on its folding, in the dynamics of interaction with the phenomena and molecules in the environment, there is a fundamental connection between the information it carries and the structure it takes once folded. Modelling such a dynamic structure allows us to understand the function of the protein. The isomorphism between the target-structure and the simulated or predicted structure is crucial to study, manipulate, and explore actual and possible functions of proteins. In so far as we need models to offer information about the target, the directionality of representation is then from model to target. It is indeed the asymmetry of the M-T relation that assures the accessibility condition (AC) that accurate models accommodate.

The isomorphic picture of the representational relation between the AF models and their T is one to take at face value if we want to develop a philosophical account of a breakthrough scientific advance such as AlphaFold.

7.8. Representational Accuracy through General and Local Scientific Understanding

In this conclusive paragraph of the chapter, I present a link between SU and representation accuracy, as discussed earlier. I defend the idea that SU requires representational accuracy.

AF models are used together with experimental methodologies to foster understanding about proteins (Laurents, 2022; Tourlet, Radjasandirane, Diharce and Brevern, 2023; Vallejos-Baccelliere and Vecchi, 2024). Even if AF models cannot *alone* allow us to gain new explanatory information about proteins' structures and mechanisms, they are now one of the most powerful tool (named method of the year in 2021), together with NMR spectroscopy, they are useful model to obtain understanding through representation. According to Laurents (2022: 1) these two methods, NMR spectroscopy and AlphaFold2, «can collaborate to advance our comprehension of proteins». He asserts that the predictions of AF can be used as structural models in a direct way or in a indirect way as tools for experimental structure determination, using other methods such as X-

ray crystallography, CryoEM or NMR spectroscopy. The scientific community is well aware of the main limits of these models. Nevertheless, we can assert that AF models play an assistive role in achieving scientific understanding of the proteins in question, although we cannot have complete explanatory information. This is due to the representational understanding AF models foster. In particular, in the case of protein electrostatics, AF plays an important role, together with other method to achieve advancements in SU (specific, local SU dedicated to certain, specific proteins).

We can distinguish general SU, which is the SU containing more theoretical advancements and implementations, i.e. historical case in physics, from the general relativity, to Feynman's partons and also the experimental detection of gravitational waves; and local SU, which is the experimental knowledge gained from the analysis, manipulation, prediction of the model.

I submit the for *general* SU, the analysis of understanding given by De Regt and Khalifa, with some adjustments suit well, while for the *local* SU more work is to be done, but I offer the analysis of ML models such as AF models and language models (which I will discuss in the subsequent chapter), as cases in which researchers gain local SU from the models. On the one hand, in general SU, the explanatory role of theories and models have been already discussed in the first and second chapters. On the other hand, De Regt and Khalifa have not told the whole story about the explanatory role of models. In the Contextual Theory of Understanding and in the EKS

model of understanding, the explanatory role of models is subsidiary to the intelligibility of theories in the first case, and the grasping of an explanation (leading to scientific knowledge) in the second case. In local SU, the explanatory role of models is paratactic to the understanding procedures, in a way that their explanatory information may not contribute *prima facie* with the model *alone* to implement the SU of the phenomenon, but thanks to the representational accuracy of the model structure, as in the case of AF, the models can enable local scientific understanding, specific to the object of study. The case of the use of AF as assistive technology together with other methods to gain SU is exemplar of this idea:

> Protein electrostatics is another important area where AlphaFold 2 has been combined with another method to achieve advances. Alone, AlphaFold 2 does not predict the pKa or charged state of the titratable residues like Asp, His or Glu. This information is very important for assessing electrostatic interactions, solubility and binding to macromolecules, substrates and drugs, but experimental pKa measurements are generally laborious (Laurents et al., 2003). Fortunately, a rather successful empirical method for estimating pKas is available called PropKa (Li et al., 2005); it uses a protein 3D structure and takes into account factors like burial, which tends to favor the neutral state and the proximity of other charged groups to calculate approximate pKa values. Thanks to AlphaFold 2, the availability of accurate 3D structures has now enabled the complete calculation of all titratable residues in the whole human proteome (Chen et al., 2022) (Laurents, 2022: 6).

It is indeed the case of protein electrostatics that Laurents (2022), after showing its use associated with NMR spectroscopy and other experimental approaches for structure determination, affirms that AF, combined with such experimental methods can «advance our understanding of protein electrostatics and forecast protein complexes. However, it does not predict rare conformations, the impact of post translational modifications (PTM), ligand binding or partially structured zones in intrinsically disordered proteins (IDPs). Fortunately, these shortcomings of AlphaFold 2 are strengths of NMR spectroscopy. In the future, results from NMR spectroscopy and other experimental methods could pave the way for future ML methods able to predict sparsely populated conformations, the effects of PTM, small molecule binding and preferred conformations in IDPs» (Laurents, 2022: 10). So, we have a general SU in which the accounts presented holds. Moreover, there is a local SU in which the explanatory role of models, albeit they do not posses the same explanatory informational core they have in cases for general SU, allows researchers to gain local SU, which may not lead to discover regularities and form complete *explanantia* of a complex phenomenon manifold such as protein folding, but improve the experimental inquiry that achieves SU when the models satisfy the criteria of representational accuracy defined above. I distinguish explanatory understanding, which is the main characterization of SU given by De Regt and Khalifa, from representational understanding, which is the characterization of SU coming from the experimental

practice of, for example, proteins structure determination *via* prediction, as showed by the research team of DeepMind and the experimental proposals of working with different methodologies together to achieve even more SU defined as representational understanding.

Note that representational understanding (RU) definition has many things in common with the analysis of intelligibility by De Regt. Visualizability and intelligibility of representational models are the key characteristics to be used in order to establish the link between the model and the target-system. It is nonsense to deny the cognitive role of scientists skill in this endeavour, but I submit that the using scientific skills to understand without looking at establishing a link between the models used and the target-system means to be left empty-handed, while looking at the link between the models and reality without scientific skills is to go blind. With this Kantian flavour, the definition of general SU and local SU can be stated as follows:

- *General SU*: general SU is achieved thanks to the knowledge of a set of explanations concerning *p*.
- *Local SU*: local SU is achieved thanks to the knowledge of a set of representations concerning *p*.

Another point to be noted is that the local SU is not an example of a renewed attempt of Lipton's proposal (Lipton, 2009) about

understanding without explanations, but is a useful distinction to draw from this study of specific Deep-learning models the main difference between the received view of scientific understanding, according to which it is explanatory, and a different kind of SU that concerns understanding through descriptions. It will be the focus of Chapter 9. Before we go there, I want to propose another case study, with which I argue in favour of a connectivist interpretation of some language models. AlphaFold and language models share both the property providing descriptive scientific understanding of the phenomena they model.

Chapter 8:
Scientific Understanding and Language Models

As in the previous chapter I have focused on bioinformatic models of protein folding prediction, now the objects of analysis in this chapter will be some examples of language models based on DLMs. In the next pages, I reconstruct the complex path of the private language argument (PLA). Then, I discuss connectionist language models and introduce notions about NPL systems' architecture. An overview of this kind of model is helpful to introduce the work of Lowney, Levy, Meroney and Gayler (2020). They submit that their model can respond to the issues raised by Wittgenstein in the notorious Private Language Argument (PLA). This argument unexpectedly turned out to be relevant not only for the philosophy of language but also for NLP and LAM modellers. I will describe the language game concept in NLP, how it is embedded, and its role in inductive systems development. This central concept in Wittgenstein's work is relevant

to describe the role of context in understanding the meanings of words . Moreover, I present the Wittgensteinian main concepts in play in the connectionist paradigm. I argue that the connectionist theoretical framework can better catch the dependency of word meaning on context. There seems to be a correlation between Wittgenstein's invitation to look, an invitation to dismiss the aim of theorizing about languages, and the absence of theory-ladenness in deep learning technologies involved in NLP. In conclusion, I criticize Lowney and colleagues' claim, whose model does not successfully capture Wittgenstein's Beetle in the Box case. Moreover, I argue that even if we can distinguish a strong and a weak definition of a private language, Wittgenstein's argument also holds for deep-learning models, and his worries are still a good guide for NLP developers.

8.1. Understanding Language and Language Models

Developers of AI systems have been struggling in figuring out how to embed in artificial platforms the specific human ability of understanding language and engaging in verbal communication with other agents. Large Language Models are become well-known objects also to laypeople, thanks to the distribution of open access chatboxes[37] based on transformer algorithms, such as ChatGPT and

[37] See on the ability (to pass the tests for undergraduate admissions), limits and features of such language models Giunti, Garavaglia, Giuntini, Pinna and Sergioli (2023).

Gemini. The great deal of attention nurtured by AI developers to understanding is discussed also by Mitchell (2019). Language understanding and the lexical competence[38] are main themes in the philosophy of language since the dawn of analysis between the XIX and XX Century.

In particular, I think is it helpful to a foundational epistemological study of deep learning language models, and the scientific understanding they provide, to start with one of the most debated theme in the philosophy of language of the last century, which has attracted also the attention of developers in computational linguistics, namely the private language argument. It is at the core of ontological and epistemological issues relevant also to this study.

Wittgenstein's ideas are a common ground for developers of Natural Language Processing (NLP) systems and linguists working on Language Acquisition and Mastery (LAM) models (Mills 1993; Lowney, Levy, Meroney and Gayler 2020; Skelac and Jandrić 2020). In recent years, we have witnessed a fast development of NLP systems capable of performing tasks as never before. NLP and LAM have been implemented based on deep learning neural networks, which learn concepts representation from rough data, but are nonetheless very effective in tasks such as question answering, textual entailment, and translation (Devlin et al. 2019; Kitaev, Cao and Klein 2019; Wang et al. 2019). In this chapter, I will debate some Wittgensteinian concepts that impact the architectures of many NLP

[38] See for example Marconi (1997).

deep-learning systems. I will focus in particular on the attempt to build a specific kind of architecture to model a private language. The discussion, I think, helps extract philosophical assumptions leading the research and development of AI systems capable of language modelling. In this chapter, I will address some of the key features of NLP systems[39] used for word embedding and one proposal to manipulate through a neural network a form of private language (Lawney et al. 2020). These modelling attempt, actually highly effective in performances, relate to the connectionist framework, that holds that «connectionist models provide a new paradigm for understanding how information might be represented in the brain» (Buckner and Garson, 2019). So, this could be also a relevant discussion about the status of the representation of the brain and its functions, theme I do not address here, but this chapter can contribute to widening the debate.

8.2. The Private Language Argument

In this section, I present the PLA and characterize it as a language game. The PLA is one of the most famous contributions of the later Wittgenstein. According to Fogelin (1976, 153), PLA is Wittgenstein's most debated argument. The debate sparked by PLA,

[39] For a discussion about LAM models, see Poveda and Vellido (2006). In these pages, the focus will be on the only NLP models.

already broad in the 70s, even increased over subsequent years and is still one of the most discussed aspects of Wittgenstein's philosophy to date. This issue raises a seduction that connects many crucial turning points in the history of philosophy. It begins with Gorgia's nihilism about the possible connection between language and the world. It continues with Galilei's discussion about the secondary quality of objects in *Il Saggiatore* (1623). Moreover, of course, it also touches on Wittgenstein's work. According to Galilei[40] , The doctrine of secondary qualities (SQs) differs from that of primary accidents, such as size, shape, motion, rest, and location. Galileo claimed that primary accidents exist in external bodies to which we attribute them. We give public ontology to the primary accidents in accordance with the use of a public language. On the other hand, we assign a private ontology to the secondary quality since these qualities «appear to exist in the objects we perceive around us and actually reside only in us» (Fisher 2009, 93). If the SQs reside exclusively in us, they can be considered private objects in a Wittgensteinian sense. Wittgenstein critiques the concept of a private ontology, offering an alternative perspective on how language operates in relation to ourselves and the world. In PI §256 and §258–60 Wittgenstein claims that a private language is impossible in which meanings are derived by ostensive internal relation to private objects. The conclusion is that a meaningful discussion about private sensation is impossible.

[40] See Galilei (1623, 28).

The expressions of individual sensations seem to imply privacy of language; this is the "primitive seduction of the private language" (Fogelin 1976, 156). Furthermore, Wittgenstein in the *Philosophical Investigations* (PI) offers two main scenarios: the first is the example of the diary of the occurrence of a private sensation (PI, §258), and the other is the monster fame case of the beetle in the box (PI, §293). The argument sketch is presented in PI §243: «The words of this [private] language are to refer to what only the speaker can know – to his immediate private sensations. So another person cannot understand the language». His attack on the idea and the possibility of a private language is contained in the passages of PI §§244– 271, even if we need to consider the broader context of §§243– 315 to follow the main ramifications of the PLA. It is not here the place to disentangle the exegetical complexity and the interpretative genealogy of the PLA. It is highly controversial whether there is, or there at least could be, a specific private language argument (PLA) to be found in Wittgenstein. Someone claims to find some prolegomenon to it in the lines of the *Tractatus Logico-Philosophicus* (TLP)[41]. Someone else sees it as a specific dialectic move of post-Tractatus[42] Wittgenstein. It is not here the place to explore a synoptic synthesis of the positions and interpretations found in the literature. However, it will be enough to cite the main four positions: orthodox,

[41] See Diamond (2000, 283).
[42] See Hacker (1984).

Kripkean, substantial and resolute. While the orthodox way[43] claims that the PLA is a reductio ad absurdum argument, Kripke[44] argues that the PLA is the consequence of the discussion on rule-following. Moreover, according to the substantial way , PLA is impossible. The resolute way instead defends the idea that PLA is possible, and that it is nonsensical to limit the possibilities of language. The first three perspectives agree that creating a private language is not possible. The first perspective views it as a deductive argument, while Kripke highlights the skeptical paradox when someone tries to develop a private language. Finally, the substantial perspective interprets Wittgenstein's work literally. The resolute reading rejects the idea that we cannot achieve a private language. In particular, the resolute readers argue that language has unlimited creative power. In fact, according to Mulhall (2007, 18), the idea of limitation is simply nonsense: no sense can be given to the idea of a philosophically substantial private language.

From §134 to §242 of PI, privacy has its central spot as a philosophical issue concerning private objects. In this chapter, I will focus on one of the famous passages: the beetle in the box. According to Stern (2011, 255): «Wittgenstein's treatment of private

[43] Norman Malcom is one of the founders of the PLA (see Malcom 1954) and one of the defenders of what lately was called the *orthodox view* of PLA. That is the idea that the paragraphs §§244–71 contain embedded an argument in the form of *reductio ad absurdum*.

[44] Kripke argued that Wittgenstein introduced a new skeptical problem to which he gave a Humean solution. According to Kripke, PLA is connected to the logical and epistemological character of following a rule. I agree instead with Hacker's interpretation of the PLA (Hacker 1984 and 2001).

language has received more attention than any other aspect of his philosophy». The pages about the Beetle in the box became famous in philosophical literature, as much as the image of Plato's cave in *The Republic* (Stern 2007), and it is even the protagonist of a title of recent philosophical text by Martin Cohen *Wittgenstein Beetle and Other Classical Thought Experiments* (2005). Let us consider the argument on its own merits, regardless of the specific Wittgensteinian style in which it was presented. Although this style is deeply rooted in his dialectic, many interpretations have emerged over the last few decades that have depicted different forms and objectives of the argument. According to Hacker, the PLA is one of «the most original and significant philosophical reflections of the twentieth century. If the line of argument pursued in them is valid, their implications, both within philosophy and without, are considerable. Modern philosophical logic, theoretical linguistics, as well as branches of empirical psychology, would stand in need of re-evaluation» (Hacker 2001, 209). Moreover, perhaps we can also add branches of artificial intelligence today. Lowney, Levy, Meroney and Gayler (2020) intentionally selected the beetle in the box as a case study to demonstrate the integration of contextual information into a language model, even if it is sourced from a private ontology.

After presenting the notion of private language, I will now describe Wittgenstein's beetle in the box case, one of the most well-known PLA scenarios:

If I say of myself that it is only from my own case that I know what the word "pain" means - must I not say the same of other people too? And how can I generalize the one case so irresponsibly?
Now someone tells me that he knows what pain is only from his own case! – Suppose everyone had a box with something in it: we call it a "beetle". No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at his beetle. – Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. – But suppose the word "beetle" had a use in these people's language? – If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a something: for the box might even be empty. – No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.
That is to say: if we construe the grammar of the expression of sensation on the model of 'object and designation' the object drops out of consideration as irrelevant. (PI, §293)

The private object to which Wittgenstein refers is not a simple invention. The beetle is there – it is an issue of private experience and how to address it. It is an example of a private experience. It could be a private sensation, a private definition, or a private object (Candlish 2004). In the background, there is a reference to the Russellian characterization of private language. Wittgenstein does not explicitly refer to the philosophy of logical atomism by Russell, but this lecture has likely been one of the targets addressing the PLA. Wittgenstein's target in PI could also be the metaphysics (a piece of plain nonsense?) involved in the Russellian logical atomism:

The results of philosophy are the discovery of some piece of plain nonsense and of bumps that the understanding has got by running up against the limits of language. They – these bumps – make us see the value of that discovery (PI, §119).

I believe that §293 of Wittgenstein's works can be interpreted as a scenario defining the limit for language games. Wittgenstein describes language in the form of language games, and I think it is possible to argue that there is a specific limit to language and language games. The Private Language Argument identifies this limit, precisely the example of the beetle in the box case. In fact, he says:

The thing [beetle] in the box has no place in the language-game at all; not even as a something: for the box might even be empty". The "thing in the box", or perhaps the same box being empty, falls outside the boundaries of the context in which we can find the meaning of the words we use. Moreover, the fil-rouge tying privacy to the limit of language games we also find in PI:

As things are I can, for example, invent a game that is never played by anyone.—But would the following be possible too: mankind has never played any games; once, however, someone invented a game—which no one ever played? (PI, § 204).

The PLA aims not to limit the possibility of expressing one's autobiographical sketch or report. The game is extensive in that

regard. The PLA issues do not impact autobiographical literature and reports. Writers have access to their interiority, and with creative acuity, they shape their private sensations, passions, and feelings, which can be shared with the public. Even though from Caesar to Valéry, Pessoa and Gide, some of the best writers of autobiographical records, it could be inferred that reaching the autobiographical truth is impossible unless the paper does not wrinkle or burn from the touch of a pen on fire (Poe 1968), it does not mean that PLA issues play a role in it. To illustrate the boundless nature of literary creativity, consider Bernardo Soares' fictional autobiography in The Book of Disquiet by Pessoa, published posthumously in 1982. In writing an autobiography, the author has the freedom to present a realistic account of their life, such as Caesar's, or a poetic diary that reflects their personal experiences and emotions, like Pessoa's during the dictatorship in Portugal. Regardless of the approach taken, the autobiography serves as a private record of the writer's thoughts, feelings and creative desires. The autobiography case demonstrates that the PLA does not intend to deny the existence of what are commonly known as private experiences that individuals have in various contexts. This clarification assists in identifying the primary targets of the PLA, which are the ontological and semantic aspects of private objects.

Taking a synoptic view of the Philosophical Investigations, we can see that the PLA plays a central role in the overall dialectic of

Wittgenstein's work. The discussions and claims that Wittgenstein presents throughout the book lead to the PLA, which is a complex argument that aims to summarize the previous discussions. Wittgenstein's dictum about language games, meaning as use, family resemblance, context, and the tenets of connectionism all have similarities, and the powerful and flexible concepts discussed in PI lead to PLA being the primary hotspot of the discussion. Therefore, using PLA to understand the connection between Wittgenstein's work and connectionism is more relevant, rather than succumbing to simplistic interpretations. In the upcoming section, I will examine the connection between Wittgenstein's concepts and connectionism. I will specifically concentrate on two major NLP models, one used for word embedding and the other used to simulate the beetle in the box scenario.

8.3. Connectionist Language Models in NLP

What is a proposition? This is the central question that TLP, PI, and NLP connectionist methodologies aim to answer. Wittgenstein writes in the Notebook, «My whole task consists in explaining the nature of the proposition» (T, 22.1.15). From this line, we must begin to trace back to the work of Wittgenstein to enlighten the contemporary exercise of connectionist methodologies for natural language. Wittgenstein's main objective was to explain the nature of

propositions in his work. However, it is essential to note that the nature or essence of a proposition should not be conceived as a platonic feature of an item existing in a detached realm of beings, happenings, and words and statements we express. Instead, the nature of a proposition is interwoven in the use of words in our language. This is why the first connection point between Wittgenstein's philosophy and connectionism can be found in Notebook and TPL, even before PI.

The first connectionist NLP method to be studied under Wittgensteinian light is Word2Vec[45]. It is a group of models based on neural network systems that produce word embedding. It is one of the first Machine Learning method used to represent the words as vector, which now seems to be overdated due to the development of transformer models, i.e. ChatGPT and Bard.

The notion of Machine Learning (ML) concerns an algorithmic process which generates an estimator that, for given input elements in a data set, values a scoring function defined over the set of output elements. The estimator is represented as:

$$f : x_i \rightarrow p_i$$

The input elements are $\{x_i\}_{i \in N} \subset \Omega_X$; and the output elements are *y-targets* $y_i \in \Omega_Y$. Usually, to parametrize f is used a set of

[45] One of the leading researchers which implemented firstly Word2Vec is Tomáš Mikolov, who introduced this technique in NLP in Mikolov, Chen, Corrado and Dean (2013).

parameters which establishes a family of estimators for the given estimation. Training an ML estimator means to optimize the estimator using the parameter values, in order to fit the data at best according to a prescribed loss function. The process to find the optimal estimator for a given training set is to train the model.

In particular, Word2Vec models produce word embedding, which is a process in which semantic structures, such as words, phrases, or similar entities from a specific vocabulary, are mapped to and mathematically modelled as Euclidean vectors of real numbers. It has a variety of applications, and it is helpful to generate text similarity, sentiment analysis, and recommendation systems. The system deploys vectorial distribution to assign a specific value to a word analysed in a context, a specific corpus. It will be likely to find in the vectorial space the word “cat” near “dog, pet, kitty, purr, paws, meow”, with a value far distant from a word that could be defined as an alcoholic drink , which defines the surrounding of, say, “wine” and “beer”. Word2Vec can utilize either of two model architectures to produce a distributed representation of words. The representation of words defines the collocation of words and their interlinguistic connections. The two models in play are continuous bag-of-words (CBOW) and continuous Skip-gram. The CBOW model predicts the current word from a window of surrounding context words. The order of context words does not influence the prediction, and this is the bag-of-words assumption. The Skip-gram model is the reverse. It

uses the current word to predict the surrounding window of context words (Mikolov, Chen, Corrado and Dean 2013).

The CBOW model is similar to a feed forward neural network. It aims to predict the current word from an output set of context words. If we input "The beetle is in the box", choose the target word "beetle" and our context words to be ["The", "is", "in", "the", "box"], this model will deploy the distributed representation of context words to predict the target word.

Instead, skip-gram is a simple neural network with one hidden layer trained to predict the probability of a given word being a context word when given a specific input word. It works as the reverse of CBOW. The Skip-gram model takes the current word and predicts the words before and after it to form its context. Given some corpus, the starting move is to select a target word over a rolling window. The researchers use pairwise combinations of the target word and all other words in the window to have a set of training data. After the training, the model assigns the probability of a word to be a context word for the given target. If we take the corpus "The beetle is in the box" and we select the target word "beetle" in a rolling windows of, say 3 words ["The", "beetle", "is"], the model will predict the probability of "The" and "is" before and after the target word "beetle".

We can appreciate how the notion of context[46] is crucial in such an NLP system. When analysing a corpus of texts, it is important to consider the context in which the language is used. This includes both the collocation of words and the extra-linguistic practices that shape our language. From Frege to Wittgenstein and modern linguistics, it is clear that both linguistic and contextual features are essential in forming the meaning of words in our language. I will dig a little into the notion of context in the next paragraph, but for now, it is important to highlight the limitations of NLP systems in relation to the context. In the Word2Vec system, every word is assigned a unique vector which codifies all its collocations and thus represents its meaning. Consequently, if two words are such that there is a context in which one of them cannot be substituted with the other, their Word2Vec vectors will be expectedly different. Another limitation concerns cases of synonymy relative to a context. Word2Vec does not operate with the notion of meaning in a particular context. Instead, it identifies the meaning of a word with a list of contexts conceived as collocations of words. An example could be run taking some statements containing the most polysemous words, such as “run”, “go” or “set”. The system will be struggling to predict the definition of the target word, that could be the same in different contexts and have different meanings which cannot be captured by the NLP models.

[46] The attempt to formalize contextual information and reduce it computationally is the crucial point of NLP systems developers; see Brézillon, Turner and Penco (2017).

The second model to be scrutinized is the VSA, which Lowney and colleagues (2020) used to model the beetle in the box case . VSA stands for Vector Symbolic Architecture; that is a connectionist model using high-dimensional vectors to encode systematic and compositional information as distributed representations (Kanerva, 1994; Plate, 2003; Rasmussen and Eliasmith, 2011). VSA family of models follows the connectionist framework of Smolensky extending it into high-dimensional vector space. (Lowney, Levy, Meroney and Gayler 2020: 652) set up a formalism comprising three operations on vectors: multiplication, addiction and permutation. According to them «VSA provides a principled connectionist alternative to classical symbolic systems (predicate calculus, graph theory) for encoding and manipulating a variety of useful structures». In fact, they suggest that «The biggest advantage of VSA representations over other connectionist approaches is that a single association (or set of associations) can be quickly recovered from a set (or larger set) of associations in a time that is independent of the number of associations». In that way, «VSA thus answers the scalability problem raised by classicists with regard to biologically plausible real-time processing» (Lowney, Levy, Meroney and Gayler 2020, 654). They choose this kind of model to capture statements similar to those of the beetle-case in PI. Their choice relies on the fact that «VSAs use multidimensional vectors and numerical weights, randomly assigned at the most basic level, in the actual processing of the networks constructed». The flexibility they attribute to the model

is based also on the fact, which is actually the Wittgensteinian tenet against the ostensive relation to private objects, that «There is no one-to-one correspondence to an entity or item for representation. A symbol is represented in signs/vectors that are distributed across a vector space, and operations with symbols, in turn, use these distributed representations to establish proximity relations that model thought and language use» (Lowney, Levy, Meroney and Gayler 2020, 654). To recognize the meaning of a word as a symbol that does not have an ostensive and one-to-one «correspondence to an entity or item for representation» is specifically to rely on the idea that «for Wittgenstein there is not typically an atomic content or correspondence that one can point to in order to explicate the meaning of a term» (Lowney, Levy, Meroney and Gayler 2020, 654). The meaning of a word, a symbol, is a product of «a complicated network of similarities overlapping and criss-crossing» (PI, §66; see also Mills 2003: 139). Following Goldfarb (1997), Strawson (1954), and Hintikka & Hintikka (1986), Lowney and the other researchers agree with the Hintikkas' way of thinking, who believe that «Wittgenstein was not denying the possibility to referring to sensations nor a private language outright» (Lowney, Levy, Meroney and Gayler 2020, 659). They use connectionism to shape a formalism in which Wittgensteinian assumptions about the nature of language are satisfied, namely that the VSA can capture some language features without assuming the connection with objects for meaning. However, the limitation of the beetle case remains fixed if we model

it with a neural network model as VSA. Connectionist models cannot explain language and its meanings insofar as – as Wittgenstein stressed – we cannot give theories of language, but only descriptions. Connectionist models, as mind models, can «help guide inquiry into the workings of the phenomena and can dispel some misconceptions, but as close as it may come to analogically portraying some important features, it should not be mistaken for the only or the actual way that language works» (Lowney, Levy, Meroney and Gayler 2020, 668). Perhaps it is better to say that the VSA proposed to model the beetle case does not resolve the beetle puzzle, even if it models the case following the line of the Hintikkas' and Hacker's interpretation, according to which there is no literal claim against the possibility of using the language to talk about private objects, as private sensations. Still, it is possible to talk about these private items using a language made by public meanings construed through interactions in extra-linguistic contexts. These contexts are not yet encoded in systems such as Word2Vec or the Smolensky vectors. With these corrections to Lowney and colleagues' proposal, I agree with their conclusion that «by respecting Wittgenstein's insights and providing a VSA account that displays linguistic compositionality, integrates soft symbols, and develops analogical structures that can be systematic and advance productively, we have shown how twenty-first-century connectionism can address what appeared to be limitations in the functionality of its operation, limitations in learning, and limitations

in biological plausibility that might have thwarted connectionism's ability to be a better mind-model for language and cognitive science» (Lowney, Levy, Meroney and Gayler 2020, 668). In this section, I have presented Word2Vec and the VSA proposed by Lowney and colleagues (2020). The two systems have underlying philosophical assumptions that were developed by Wittgenstein. Consequently, they show how Wittgenstein's ideas are deeply embedded in the deep-learning NLP models, and how his ideas are integral to the breakthrough of AI language models. In the following section, we will explore how Wittgenstein's conceptual tools are closely linked with the main connectionist principles.

8.4. Wittgenstein and Connectionism

Between the 60s and the mid-90s, two main approaches concerning LAM were predominant: the connectionist and the symbols and rule approach. There are two main approaches to explaining cognition. The first considers it as an ability that can be understood through a neural network. The second approach, the symbolic or rule-based approach, views the mind as an information processing system similar to a computer. In this view, language is seen as a collection of symbols that are governed by rules, similar to how programming languages are used in computers. At the time the symbols and rule approach was predominant, Mills (1993) stated

some important similarities between Wittgenstein's later philosophy and connectionism. Building on this view, Lowney, Levy, Meroney and Gayler (2020) specify the affinity between Smolensky's (1991) distributed representation connectionist approach and Wittgenstein's main philosophical notions, such as symbol constitution, language-games, family resemblance, rule-following, logic, and language learning. According to them, there is a possible path connecting Wittgenstein and connectionism.

The last decades have witnessed an upswing in the debate between the classical approach and connectionism, and many scholars have made an important effort to disentangle the Wittgensteinian root of connectionism and deep-learning models. According to Stern (1991), Mills (1993), Goldstein and Slater (1998), and Elman (2014) connectionism provides an understanding of mind and language use that can be traced back to the so-called later Wittgenstein. Connectionism states that cognition is an emergent property relying on associations and activation of patterns following parallel processing. Dating to Skinner's work, Chomsky, Fodor, and Pinker proposed an alternative to connectionism through the renewed classical symbolic approach. Now, we could recast connectionism as a cognitive science movement working on deep-learning models. Recently, it has been noticed the importance of Wittgenstein's work for both Natural Language Processing systems and connectionist theoretical frameworks, in particular deep-learning neural networks. NPL and LAM models rely on deep-learning neural

networks in both cases. The use of deep-learning models to analyse the Wittgensteinian concepts in NPL and LAM offers a perspective that helps clarify the underlying assumptions in current connectionism ideas.

Connectionism, as an alternative to classical approaches and as a family of concepts and methodologies, is much more effective than symbolic AI. According to Mitchell (2019: 38): «A symbolic AI program's knowledge consists of words of phrases (the "symbols"), typically understandable to a human, along with rules by which the program can combine and process these symbols in order to perform its assigned task». It was thanks to the back-propagation, a general learning algorithm, towards which Misnky and Papert (1969) were sceptical, that multilayer neural networks played a crucial role in the foundation of modern AI and led to the rise of Machine Learning (ML). Back in the 1980s, at the University of California, San Diego, there was the most known team working on neural networks led by psychologists David Rumelhart and James McClelland. This team is well-known today for the writing of Parallel Distributed Processing (1986), recognized as the "bible of connectionism" (Mitchell 2019: 39). According to them, the symbolic systems such as those favoured by Minsky and Papert (Rumelhart, McClelland, and the PDP Research Group 1986: 113) would fail to catch the humanlike abilities such as perceiving objects, understanding language and retrieving information from memory. According to Mitchell, «What we now call neural networks were then generally referred to as

connectionist networks, where the term connectionist refers to the idea that knowledge in these networks resides in weighted connections between units» (Mitchell 2019: 39).

Connectionism has been defined as a neurologically inspired architecture: «connectionist nodes and networks were said to model neurons and the synapses created through the connection of axons and dendrites. Connectionism thus showed promise for unifying brain biology and perception with higher cognitive activities such as thought and language» (Lowney, Levy, Meroney and Gayler 2020, 645). Connectionism implies consequences about how we understand the language and the mind that early Wittgenstein figured out in his later work (Stern 1991; Mills 1993; Dror and Dascal 1997; Goldstein and Slater 1998; Elman 2014). In fact, according to Mills (1994, 145) connectionist training is very similar to the network experiencing the proper use of the words, in particular language games. Connectionist models rely on both the notion of language games and contextual features to develop powerful tools.

Wittgenstein's transition from TPL, through *The Blue and Brown books* (BB), to PI, is fundamental for the conceptual redefinition of these concepts. In the pages of BB, language games are conceived as primitive forms of language complete in themselves, yet imagined as evolving in changed circumstances into new and more complex ones. Moreover, in BB, a language games is defined as a *Satzsystem*. The language game is thought of in analogy with an axiomatic system. It involves a large scale semantic holism. It will be in PI that the

concept of *Sprachspiel* updates the preview developments. Here the language games are analogous to games in which the meaning of words is inextricably tied to the speakers' non-linguistic practice - «the whole, consisting of language and actions into which it is woven» (PI, §7). Here emerges the force of the universe of practice and actions, and especially the crucial role of the public context. The study of language in the 1950s is disciplined by linguistics, and Firth (Church 2007) is one of the leading figures that pinpointed the role of context. He was convinced that the complete meaning of a word is always contextual, and no study of meaning apart from context can be taken seriously and coined the notion of collocation. Collocation is «quite simply the mere word accompaniment, the other word-material is which [the word is] most commonly or most characteristically embedded» (Firth 1968, 180). According to him, context-collocation is a part of the word's meaning (Firth 1969). His fame followed the slogan, «You shall know a word by the company it keeps» (Firth 1969, 179). As we have seen in Word2Vec and in VSA for the beetle, the context and the collocation are crucial issues for the connectionist models. Moreover, it is again the context to be the hinge of the Wittgensteinian concept of meaning-as-use. This is a contextual concept. Meaning is public and accessible insofar as words are contextualised (in a specific language game).

Given that understanding and mastery of a language is a practice and that the meaning of the words is the use we make of them, as Wittgenstein stated in the Philosophical Investigations , it follows

that the contexts of specific situations in which we learn, use, and understand the language, is a key item in the analysis of NLP and LAM. According to Wittgenstein, the contexts in question have public features and are crucial for his description of language games. Hence it is not possible to develop linguistic abilities in a private context, such as those described in the case of the beetle in the box. The conceptual role of contexts is crucial for both NLP and LAM. AS concerns NLP, looking at the shallow neural network of Word2vec, a word vector representation (WVR) applied in a translation model built by Google, the concept of context is of recognized importance given the output the network must give. Moreover, Lowney, Levy, Meroney and Gayler (2020) showed how Vector Symbolic Architectures (VSA) can resolve some limitations encountered by the previous connectionist approach concerning the private language. Even if we distinguish a strong and a weak sense of the argument of the private language, the concept of meaning as occurring within a public context is assumed by NLP. In conclusion, WVR and VSA neural networks are built with different purposes, the former for translation, the latter to explain language acquisition and mastery, but both are grounded on the assumption that the meaning of linguistic signs could change according to the contexts in which they are used.

To conclude we could say that Wittgenstein paved the path for the computational study of language: that is a connectionist analysis of language. We have seen that connectionism employs Wittgenstein

ideas and it is coherent with the broad picture of philosophy: connectionism offers a way of describing the language – it is coherent with Wittgenstein's suggestion to stop the ambition of theorizing about meaning and therefore to describe. From the symbols and rules approach to connectionist models we see the evolution of Wittgenstein's work: «While the symbols and rules approach might provide an emergent level of description, we see a disjunction, though not a sharp one, at the level of symbols (as encapsulated representations or dynamic attractors) and then at the level of roles (in which rule-like features come from connectionist ability to build analogical structures). These transitions suggest a more fluid picture of language than the classical approach allows and suggest an emergent rather than a reductionist or implementational account» (Lowney, Levy, Meroney and Gayler 2020, 666). Before moving on, let us summarize the relationship between Wittgenstein's ideas and connectionist models, such as Word2Vec, Smolensky models, BERT, or GPT. We claim that all of these models are computational upgrades of language games. In the following section, I critique the idea that the VSA can simulate a private language and provide an instrument to simulate the reference of private objects. The precedence of context over linguistic practices in language games remains an insurmountable barrier to those who argue for the feasibility of a private language.

8.5. The Beetle in the (Black) Box

As Mills (2003) noted, for Wittgenstein there is not typically an atomic content or correspondence that one can point to explicate a term's meaning. This is what follows from the story of the sensation diary and the story of the beetle. In these two stories, there is a relation between a human being, an interior (private) object, and the language. If we want to model this framework with computational tools and connectionist methodologies, we will have a relation between human language and a system of algorithms (running on a device). An analogy could exist between what happens in the Wittgenstein scenarios and in connectionist models. Indeed, if the connectionist model is a kind of analogue to the mind model implied in PLA, the two models share similar features. The most important one is the analogy between the inaccessibility of the private objects to the others in PLA and the inaccessibility of the black box in the algorithms of connectionist models. Due to a proprietary copyright shield, some algorithms are programmed to contain or shadow some processes within a black box (this is the case with Google search, Instagram and Netflix). In this case within the black box, the content is not private as in the beetle case. However, this could be an analogy for the beetle case and the solution proposed by Lowney, Levy, Meroney and Gayler (2020). They propose a method to hide/shadow the meaning of an interior object, shifting it from one meaning to another so that my beetle could be different from your beetle. It does

not resolve the problem highlighted by Wittgenstein. The problem of PLA cannot be solved and demarcates the limits of the language games. According to Rudin and Radin (2019, 2): "In machine learning, these black box models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions". Take all the Xs as processes within a black box, which seems to behold to the "privacy" of the machine – we cannot understand what is going on there. The analogy between Wittgenstein's scenarios and the black box property of algorithms does not want to play the role of anthropomorphizing the machines, "it makes no sense to ascribe thought or thoughtlessness, understanding, misunderstanding or failure of understanding to machines" (Hacker 2011: 34), but to highlight similar features, similar boundaries in both cases. As Rudin and Radin say, the black box hides the algorithmic processes, so how the variables are combined cannot be understood to make the final prediction. Black boxes limit our ability to understand data processing. The black box contains information we cannot access, as in the beetle case. The VSA proposal we have seen does not overcome this limitation. Lowney, Levy, Meroney and Gayler (2020: 655) recognize the inaccessibility of the private object: «Together, the iterations encountered in social contexts provide us with common language-games about, e.g. beetles, but there need be no specific entity directly represented by a VSA symbol». Again, they say that "VSA's use of

random vectors virtually guarantees that “my beetle” will be different from “your beetle”, thus mirroring the variations in the uses of the word that we each have encountered”. Their model aims to capture the shifting nature of the meaning of the word “beetle” in different contexts: «Further, as the concept or symbol BEETLE evolves in the experience of an individual, or the world “beetle” changes its use by that individual in different contexts, the random vectors that constitute the symbol itself may be “constantly changing”, but these differences, at what becomes the sub-symbolic level, do not interfere with the informative communication that can take place at the symbolic level due to the regularities in the use of the words» (Lowney, Levy, Meroney and Gayler 2020: 655). What must be specified is that it is not possible to capture the meaning of “beetle” as in the Wittgenstein scenario. What can be done with the VSA model is capturing the use of “beetle” in some language games, in a public context. Their proposal to develop an algorithm to catch some uses of BEETLE is thus a kind of language game. They propose a way to describe what we can do publicly with language. Better, they tell how we can play with a philosophical problem, namely PLA. Nevertheless, their argument works only if we take the weaker version of PLA, in which we can deal with different private objects with slightly different meanings and communicate about them to each other. Therefore, in their scenario, the privacy of the meaning disappears. Their proposal is coherent, though with the suggestion to describe the language. Indeed, Wittgenstein claims that theorizing

about language, including private language, is not a good philosophical practice.

Some scholars have worked on the importance of context for human reasoning and communication. In particular, Hollister, Gonzalez and Hollister (2017) extend the discussion about contextual reasoning in humans and how modelling it in a computer program can help to get closer to the ultimate intelligent machine: «We strongly believe that to create systems that have the full range of human-like intelligence requires imbuing them whit the ability to process context – giving the system an idea of when and where the information was previously encountered so that a solution might be found using the situational context» (Hollister, Gonzalez, and Hollister, 2017: 599). AI developers face a terrific challenge: implementing contextual intelligence in deep-learning systems. The assumption is that context precedes language game in action. The ability to understand and master a language is based on contextual experience. The context precedes the speakers' awareness of the linguistic experience. Lowney and his colleagues face a significant challenge: they want to represent a situation where we do not have any information about the beetle's owner contextual experience as a private object when using VSA. In other words, we end up with an empty dataset related to the beetle owner's contextual experience. To function effectively, VSA requires incorporating contextual information inherent to the language used by speakers. This approach validates Wittgenstein's PLA and reinforces that language is a shared

activity rather than providing further support for the philosophical stance of resolute readers.

The beetle case, as explained by Wittgenstein and Lowney and colleagues (2020), highlights the importance of context in understanding the meaning of words. In the absence of context, words lose their meaning. Context plays a crucial role in explaining the meaning of words. Word2Vec, a sophisticated word embedding system created by humans, cannot capture the nuances of open concepts. The analogy of privacy illustrates the limitations of NLP systems in grasping the meaning of words. Word2Vec and VSA are language games lacking the contextual features inherent in our daily practices and form of life.

As we have seen, deep learning models are a type of artificial intelligence that learns to perform tasks by analysing large amounts of data. These models are often used in natural language processing (NLP) applications. Deep learning models are trained on large datasets of text and code. The model learns to associate patterns in the data with specific outputs. Once a deep learning model is trained, we can use it to perform tasks on new data. Indeed, deep learning models have been very successful in many NLP tasks. However, they are not able to represent the contextual features of meaning that are essential for understanding the beetle-in-the-box case. The private language argument does raise important questions about the ability of deep learning models to understand the meaning of words and create private meanings. Overall, the private language argument is a

reminder that we need to be careful about making claims about the ability of AI to understand human language. While deep learning models have succeeded in many NLP tasks, they still have limitations.

Wittgenstein's private language argument shows that the meaning of a word is not determined by its private reference to some internal object or state of mind, but rather by its use in a particular language game. Deep learning models are not able to represent the contextual features of meaning that are essential for understanding the limitations of the beetle-in-the-box case. Therefore, deep learning models cannot be used to represent the private language argument. They represent the limit of language games and the meaning that can be derived from them.

8.6. Scientific Understanding and ML Language Models

In this chapter I have analysed some ML methods to develop NLP systems. These models are used also to foster experts understanding of language. Albeit it is one of the most debated and controversial subject in AI and also General AI, whether machine can gain understanding[47], both semantic and scientific: «While state-of-the-art AI systems have nearly equaled (and in some cases surpassed) humans on certain narrowly defined tasks, these systems all lack a

[47] See Mitchtell (2019, ch. 14) and the interview with Ellie Pavlick (Pavlus, 2024).

grasp of the rich *meanings* human bring to bear in perception, language, and reasoning. This lack of understanding is clearly revealed by the un-humanlike errors these systems can make; by their difficulties with abstracting and transferring what they have learned; by their lack of commonsense knowledge; and by their vulnerability to adversarial attacks. The barrier of meaning between AI and human-level intelligence still stands today» (Mitchell, 2019: 235-36). There are many scholars that are working on how computer could encode semantics, like Ellie Pavlick, whose area of research focuses on "grounding": «the question of whether the meaning of words depends on things that exist independently of language itself, such as sensory perceptions, social interactions, or even other thoughts» (Pavlus, 2024). The relation between things that exist independently of language itself is at the core of the discussion about the PLA had in this chapter. As we have seen, and also Pavlick stresses, the "grounding" of meaning has a relation to be studied with the extra-linguistic context. Now, we know that language models are trained entirely on text, without contextual information about the extra-linguistic world. The issue about how grounding matters to meaning has attracted the attention of linguists and philosophers for decades, and still, these problems are not only of a technical nature: «There are not only 'technical' problems [..]. Language is so huge that, to me, it feels like it encompasses everything», says Pavlick (Pavlus, 2024). So, understanding of meaning is a property under scrutiny in the domain of DLMs, specific to certain language models.

If we want to abstract, though, we do know that a higher level question is in the air when discussing such issues, and it is whether – and to what extent – we can say that in general DLMs understand, or, in a more secular way, that DLMs help us understanding the phenomena we study. On this second issue is focused the descriptive interpretation of scientific understanding secured by DLMs I will present in the next chapter.

8.7. ML Models and Black Boxes

The analogy between the case of the Wittgensteinian Beetle in the box for natural language games and the Black box for deep-learning language-game relates to the general black box property of algorithmic structure. Black box problem comes from an overlapping of different issues: opacity problem, the strangeness problem, the unpredictability problem, and the justification problem (Brożec et al, 2023). But the main interpretation of ML models is that they are implementation black boxes (Sullivan, 2022). Discussing the case of melanoma analysis with ML methods, she argues that «implementation black boxes do not get in the way of understanding phenomena in the melanoma case because the model is operating within a background existing scientific understanding. […] the level of scientific justification and background knowledge linking the appearance of moles to instances of melanoma is extensive», and

effectively it is a «leading deciding factor for medical intervention» (Sullivan, 2022: 126), and eventually chirurgical operation and biopsies. This kind of implementation according to Sullivan is then harmless. Moreover, Tamir and Shech argue that this stance supports the thesis that there may be noticeable relations between features of the phenomenon, but it is not enough to have clear links for understanding: «Certainly, background scientific knowledge can inform the kinds of features to target with an ML model[48], but in order to establish a link between the model and the target, more is needed» (Tamir and Shech, 2023: 335). The more is the properties of the system thank to which the model can represent the target-system. But how the representation obtains from the interpositions of hidden layers, this is an issue that deserves further attention.

The relation between the data elaborated through the hidden layers and the input and output of the ML models is also at the centre of the opacity problem[49]. This problem is articulated in different issues, one of which is the representation learning of hidden layers. One influential discussion about it is carried out by Bengio et al. (2013), that describe how DL models have to learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data.

[48] This is the case of the background biological and informatic and mathematical knowledge necessary (along with high expertise in ML methods) to develop such models as the AlphaFold's ones.
[49] See, for an extensive presentation of the very many kinds of opacity in AI systems, Facchini and Termine (2022).

One example of the informative hidden layer representations is, as we have seen, the words vectorization. Contemporary more complex Transformer techniques, as Tamir and Shech note (2023: 337-38), use deeper pre-training of text embedding methods, that are adapted to embed chunks of text and individual terms in the context of surrounding text in which they are written (Devlin et al., 2018; Vaswani er al., 2017). As the example of Word2Vec shows, Mikolov et al. (2013) use shallow neural networks to map individual terms to vector representations, optimized for tasks such as "fill in the blank". According to Tamir and Shech (2023: 332), this vectorized word-embedding system is not just useful for the original task, but the representations elaborated could be reused as pre-trained representations for novel text-based tasks. In fact, as we have seen, word-embeddings have been used to study and use ostensible semantic relationships, i.e. analogies, synonymy clusters, manifested by the usage patterns for practical tasks and applications. I have then shown before that we must take carefully the representation of the ostensible semantic relationships, given the limits the system exhibits under the test of the private language argument.

Chapter 9:
Descriptive and Explanatory Scientific Understanding: Towards a Pluralism of Understanding

In this chapter the difference between descriptive and explanatory understanding is presented and I argue that on the basis of this difference two ways of conceiving the scientific understanding with deep learning models (DLMs) are possible. In virtue of the characteristic of descriptive and explanatory understanding, we can argue that the scientific understanding of DLMs is important to achieve SU from them, or not. I submit that, as we have seen in the Chapter 1. for explanations. Different models are useful for different stages of the scientific research process, so the scientific practice can use explicatory models in different moments and areas. I think that the same can be told for scientific understanding. As we have seen, in Chapters 2, 3 and 5, 6, there is a distinction between explanatory and descriptive understanding, and maybe the future study of understanding will show different models and theories of it. Each

model of SU can be applied to different stages and different areas of the scientific research. From this idea arises the pluralism of scientific understanding, here developed according to the main two branches, the explanatory and the descriptive.

9.1. The Elements of Scientific Understanding Pluralism

In these pages we have seen how entangled the notion of SU is with explanation and models. There are many kinds of understanding and many ways in which the inquirers can achieve scientific understanding, also with sophisticated technological constructs as DLMs. Thanks to the discussion animated by Khalifa and colleagues (2023), the friends and frenemies of understanding are working to find a common ground to establish a third way to define scientific understanding. It is in this common space that I want to advance the main ideas of what can be call a *pluralistic* account of scientific understanding. A third way where friends and frenemies of understanding would go hand in hand. It is not a detached alternative to De Regt's and Khalifa's account, but is a proposal to integrate them in a broader – and more pluralistic – framework. In Chapter 6. we have already seen the differences between the global and local SU, here I want to build on this difference to account for descriptive scientific understanding on one side, and explanatory on the other. While, as in Chapters 5. and 6. is outlined for case studies in

bioinformatics (AlphaFold) and computational linguistics (VSA models for private language), DLMs offer a kind of understanding I define as *descriptive*, the models taken by physical sciences, as the case of ideal gas model used by De Regt and the example of Bjorken scaling used by Khalifa, are exemplary to account for *explanatory* scientific understanding. In the pursuit of understanding the complexities of the world, various disciplines offer unique lenses through which we can interpret and understand phenomena. From cognitive science to philosophy of science, diverse perspectives contribute to our evolving understanding of scientific understanding (USU). This essay explores the concept of pluralism in SU, integrating insights from scholars such as Danielsson, Sullivan, Beisbart, Räz, de Regt, Khalifa, and others. By examining the intersection of cognitive abilities, explanatory frameworks, and technological advancements, the aim is to construct a pluralist framework that embraces the multiplicity of approaches to SU.

Danielsson's assertion that our comprehension of the world remains incomplete until we fully understand the intricacies of our mind and brain resonates with Sullivan's emphasis on understanding Deep-learning models. Similarly, Beisbart and Räz (2022) advocate for considering both understanding DLMs and understanding with DLMs. This pluralistic view seeks to integrate diverse perspectives on SU, including descriptive (Johannes Findl and Javier Suárez, 2021) and explanatory conceptions, as proposed by de Regt and Khalifa.

As we have seen, De Regt choses to decouple the notion of understanding from the traditional image of explanatory understanding, in a pragmatic and liberal way; the latter is bounded to the received view of understanding, claiming that scientific understanding is deeply bound with scientific knowledge, re-establishing the link between scientific understanding and explanatory understanding.

Given that a definition of grasping involved in understanding remains unsolved, the issue of the definition of understanding as a cognitive ability that is the main concern of psychological, neuroscientific, linguistics and philosophical research, remains still open. What an agent's understanding consists of is defined as a cognitive ability which is called "grasping" that enables the subject «to draw the conclusion that *p* (or probably *p*) from the information that *q*» (Hills, 2016). This broad definition of "grasping" allows to conceive the notion of information locally or generally. From this point we can distinguish between:

- *General SU*: general SU is achieved thanks to the knowledge of a set of explanations concerning *p*.
- *Local SU*: local SU is achieved thanks to the knowledge of a set of representations concerning *p*.

A distinction like that one above is at the bottom of the thesis that we can have explanatory understanding distinct from descriptive

understanding of phenomena. The detail of this kind of difference will be given later on.

Thus, the components of a possible pluralistic account are the following:

1) *Pragmatic Approach*: de Regt's pragmatic and liberal stance decouples understanding from the traditional image of explanatory understanding, emphasizing the importance of skills and contexts.
2) *EKS Model*: The EKS model (Khalifa, 2017), with its focus on knowledge and voluntarism about realism, provides a framework for explanatory aspects of SU.
3) *Representation, Description and Understanding*: Beyond explanatory understanding, descriptive understanding, proposed by Johannes Findl and Javier Suárez, complements our understanding of phenomena without conflicting with explanatory understanding.

An overlapping area emerges where the perspectives of de Regt and Khalifa collide. While de Regt emphasizes the integration of knowledge and skills, Khalifa highlights the importance of considering both explicit propositional knowledge and implicit skills in the scientific endeavour. This necessitates a broadened conception of knowledge that encompasses various forms of understanding.

Moreover, despite the merits of de Regt's approach, challenges arise when considering cases such as AI assistive understanding, exemplified by AlphaFold models. While these models offer accurate predictions[50], they may lack a clear account of underlying explanations. Vellejos-Baccelliere and Vecchi's distinction between thermodynamic and kinetic explanations in protein folding research highlights the complexity of explanatory understanding in such contexts. There are two main explananda in Protein Folding research, the thermodynamic and the kinetic explanations.: the first concerns the factors that determine the stability of the native structure (native state stability issue); the second concerns how the protein acquires its native 3D structure starting from an unfolded state (the folding dynamic problem). Both levels of understanding proteins, the kinetic and the thermodynamic, have to be analysed at the same time, to achieve general understanding of folding phenomena: «The study of protein folding plays a crucial role in improving our understanding of protein function and of the relationship between genetics and phenotypes. In particular, understanding the thermodynamics and kinetics of the folding process is important for uncovering the mechanism behind human disorders caused by protein misfolding» (Turina, Fariselli and Capriotti, 2023: 1). This is consistent with the decoupling of understanding and explanation. As prediction becomes increasingly independent from explanation, as noted by Vallejos-

[50] See for an overview about prediction in Artificial Intelligence Bianchini (2018). For studies on the relation between prediction and understanding in scientific modelling, see Elgin (2017), Khalifa (2017) and Potochnik (2017).

Baccelliere and Vecchi (2023), the study of protein folding remains crucial for uncovering mechanisms behind human disorders caused by protein misfolding. Despite the limitations of AF models in providing standalone explanatory information, descriptive understanding achieved with assistive technologies opens avenues for exploring alternative modalities of understanding beyond conventional explanatory frameworks.

Indeed, de Regt's and Khalifa's views have come points in common; there is one overlapping area in which the desiderata of De Regt and Khalifa collide: «Perhaps the two views can be reconciled if it is acknowledge that knowledge and skills are intricately related, and that science needs both. This would require a broadened conception of knowledge, on which "knowing" includes both explicit, propositional knowledge and implicit skills» (de Regt, 2023: 30). De Regt's approach, characterized by its liberal stance on the attainment of understanding and its emphasis on skills and contexts, resonates with contemporary notions of scientific inquiry. However, a notable deficiency in his framework lies in its failure to provide a compelling account of the explanatory mechanisms inherent in understanding. This becomes particularly evident when considering cases such as AI-assisted understanding, exemplified by AlphaFold models. In such instances, understanding is derived primarily from the predictive capabilities of the model, rather than from a direct application of background theoretical knowledge. Despite the model's ability to produce realistic and intelligible outputs, its

explanatory information remains questionable. For example, within the context of AlphaFold models, let Q represent the model, P the amino-acid strain, and T the target-system, i.e., the folded protein. The pressing question arises: Does the model genuinely furnish explanatory information? This query underscores the complexities surrounding the intersection of computational modelling, theoretical knowledge, and the pursuit of scientific understanding.

In conclusion, a pluralistic approach to SU acknowledges the diversity of perspectives and methodologies inherent in scientific inquiry. By integrating insights from cognitive science, philosophy of science, and technological advancements, we can construct a more comprehensive framework that embraces the multiplicity of approaches to understanding the world around us. Through ongoing interdisciplinary dialogue and exploration, we continue to refine our understanding of SU, paving the way for new discoveries and insights into the nature of world and cosmos.

9.2. Scientific Understanding and Machine Learning Models

Here, to conclude, I want to present the main elements of the debate about understanding with and understanding of machine learning (ML) models (and Deep-learning models as a subcategory). There are two sides of the debate. On one hand Sullivan (2022) argues that the understanding *of* ML models does not impact on

understanding *with* ML models. Although we can fail when we try to understand a phenomenon through ML models. According to Sullivan, though, the failure is due to the "link uncertainty", that is when we lack evidence and knowledge about how the model and its output are linked to the phenomena studied or the target-system. On the other hand, Räz and Beisbart (2022) argues, *contra* Sullivan, that if we fail to understand a phenomenon *with* ML models, it is because we lack understanding *of* the ML models, of how they work. As they reconstruct Sullivan's argument, they submit that she refer to the understanding provided by ML models as explanatory, namely as providing how-possibly explanations. They argue then that how-possible explanation need not be an explanation (Räz and Beisbart, 2022: 11). I think that they are right, the kind of scientific understanding we can gain from ML models, and DLMs in particular, is descriptive, not explanatory.

To support this view, I follow Tamir and Shech's ideas, that argue that «while DL models do not necessarily provide understanding in the same manner as (say) map representations, learned representation layers may be leveraged to improve understanding by providing insight into how a DL models learns to organize raw input data to optimally estimate y-targets» (Tamir and Shech, 2023: 327). Let's consider an ML NLP model trained to solve the exercise "fill the gap", or "complete the sentence", supported by algorithms of word embedding, as we have already seen. According to them, the appropriate target of understanding with the aid of ML models is

closely related to how a learned distribution p_θ (Y|X) estimates the actual distribution p (Y|X) describing the phenomenon's features. They propose this definition of the target of ML (TML):

Target of ML Hypothesis (TML): the target phenomenon of understanding with ML models is the *relationship(s) of features represented by the data*.

They argue that the target phenomenon is not a particular object or sampling instance, but relationships of the properties or features found potentially in individual or multiple objects or object types. So, according to them, «if TML hypothesis is correct […] ML models help us understand relationships between features represented by the data, but expecting such models to further provide causal explanations (e.g. for why a feature is predictive) is inappropriate» (2023: 349). TML is indeed model centric, in fact the target they have in mind consists of understanding the model, and not how the model is useful to gain SU of the phenomena. This it their flow to be corrected. A correction to TML is provided by Sullivan (2023) and I agree with her implementation: «when models enable understanding of real-world phenomena (or possible real-world phenomena), the target of understanding is not the relationships of features represented by the data. Data relationships are used as a means to some further end, or, as I have argued elsewhere with Insa Lawler, models provide understanding by *inducing* explanations of

phenomena instead of the model itself being an explanation» (Sullivan, 2023: 345; Lawler and Sullivan, 2021). According to Sullivan, the ML model alone does not provide understanding, the model *induces* an explanation which is «paired with external link connecting the model to the phenomena that enables understanding» (Sullivan, 2023: 345). This is the main reason why she defends the idea that most ML models provide how-possibly explanations, and not causal ones. Eventually, ML model indicate possible causal hypotheses that only with additional research can be justified through the reduction of link uncertainty. She concludes that «restricting ML models to the model-centric view of the TML goes against the goals that many ML researchers themselves postulate, as well as keeping ML models apart from the rest of model-based science, which unnecessarily constrains our scientific toolbox» (Sullivan, 2023: 345). I argue that the main flow in this line of thinking is to take the kind of understanding we want to attribute to the outcome of the scientific research that involves DLMs as explanatory, while it is of another sort. It is indeed on the representational property of the models that we have to focus, to account for the understanding of phenomena scientists gain in the use of these models.

9.3. Representationalism and Understanding

The origin of the representationalist image of understanding has a venerable tradition. Yet the image is incomplete, insofar as the core of representationalism I defend in this pages concerns the feature of representational information useful to achieve descriptive understanding of phenomena. According to Kuorikoski: «The representationalist or intellectualist tradition understands understanding as the private possession of correct mental representation of the objects of understanding. In contrast, the pragmatist tradition, drawing on Kant, Wittgenstein, and Sellars, understands understanding as a matter of doing. In Ryle's parlance (1946), representationalist understanding amounts to knowledge that and the pragmatist understanding to knowledge how» (Kuorikoski, 2023: 218). Kuorikoski defends a pragmatist and deflationary view of understanding, which identifies explanatory understanding with the ability to draw correct counterfactual what-if inferences about the object of understanding (Kuorikoski, 2011; Kuorikoski and Ylikoski, 2015). There is an emerging consensus in philosophy of science that the distinctive feature of the explanatory knowledge embedded in understanding is its modal dimension and to understand a phenomenon is to be able to correctly situate it within a space of possibilities (Kuorikoski, 2023: 218). I agree with the view that the distinctive feature of explanatory knowledge implied in understanding is its modal dimension (modality of possibility) – and

the exploration of the space of possibilities is an image that suits well with the AI project of DeepMind to design a tool to predict the proteins' structures. AF architecture and models are designed to explore such "space of possibilities", in order to track down the approximately true structure of the target-system. Nevertheless, I disagree on the distinctiveness of the modal feature applied only to explanatory knowledge equated to explanatory understanding. So, I distinguish explanatory understanding from a kind of representational understanding, the modal dimension of which is the distinctive feature of descriptive understanding.

9.4. Descriptive and Explanatory Understanding

De Regt advances the contextual theory of SU and Khalifa the EKS model of SU, which can be conceived as capturing the features of Explanatory Scientific Understanding (ESU). In the cases of AlphaFold models and Deep-learning models of language, a different kind of understanding has been pictured, which can be called Descriptive Scientific Understanding (DSU). In making this distinction, I follow the suggestion given by Findl and Suárez (2021), according to which it is correct to distinguish between explanatory and descriptive understanding, in particular, they use the case of epidemiological statistical Covid-19 models, early developed by the Institute of Health Metrics and Evaluation (IHME). They argue that

«early epidemiological models yielded a modality of understanding that we call *descriptive understanding*, which contrasts with the so-called *explanatory understanding* which is assumed to be the main form of scientific understanding» (Findl and Suárez, 2021: 1). I will present this point in the next paragraphs, but let me now give you the two depictions of EU and DU, in order to hold already the main claims of this chapter:

- *Explanatory Scientific Understanding (ESU)*:
  The agent *A* has explanatory scientific understanding of *p* if *A* grasps a set of explanations *S*, such that *A* can represent *p* throughout the use of *S* in a theory *T* or in a model *M*.

- *Descriptive Scientific Understanding (DSU)*:
  The agent *A* has descriptive scientific understanding of *p* if *A* can appreciate a representation *R* of *p*, such that *A* can infer from *R* relevant information about *p*.

Since ESU requires the possession of an explanation for an adequate occurrence of scientific understanding, this limits the application of ESU on some scientific practices. The main limitation of ESU is that the model does fail to capture the features of SU gained with statistical models and DLMs. On the other hand, DSU offer a countermeasure to ESU limitations, in so far as the representation of *p* is descriptive. I means that it can account for a prediction of *p*. May

the prediction concern the structure of *p*, as in the case of AlphaFold models, or the evolution of *p*, as in the case of the statistical COVID-19 models, or the relevance of *p* according to many possible outcome, as the distinctive features of language models can suggest.

The main difference between ESU and DSU is that the former captures the understanding activity related to the explanatory set of information of a phenomenon, while the latter depicts the understanding activity related to a descriptive representation of a phenomenon. An example of this difference relies on the different angles a scientific community has in inquiry, i.e. the AlphaFold models that answer to the question how certain proteins fold and not why. It is the case also of the point made by the chemical biologist Shuibing Chen (Weill Cornell Medical College, New York), that in an article of *Nature news* by Sara Reardon[51], about mini-colon and brain organoids model that could shed light on cancer and other diseases, says: «In the last ten years, people spent a lot of time to develop and *understand how* [mine italics] to make organoids. But this is the time to think more about how to use» the models (Reardon, 2024). The kind of understanding involved here is explicitly understanding how, which can be conceived as a case of descriptive understanding (DSU) of a phenomenon, namely organoids. Understanding how to make organoids and understanding

[51] See Reardon (2024) for the extended article about organoids models and their use against cancer and other diseases; for an introduction to model organism, see Ankeny and Leonelli (2020); for a detailed study of organoids properties, see Picollet-D'hahan et al. (2017), and Laplane et al. (2019) for the philosophical study of organoids as exemplary reason to show the science being in need of philosophy.

how to use the organoids models are both cases of DSU, due to the features of the understanding involved, that is in both cases determined by a set of instructions to be followed to hit a specific outcome: to make a model and to use it. The use of a model can be rephrased as a set of rules and instructions in order to fulfil a function or to solve a problem.

The ESU is characterized by Findl and Suárez (2021) on the basis of the analysis made by de Regt (de Regt, 2009, 2017; de Regt and Dieks, 2005), who conceives explanations as mediators between prediction and understanding. In their account, indeed, they use the case of the early COVID-19 models to link prediction and understanding in their account of *DESC*, the descriptive understanding they defend. De Regt builds his contextual theory of SU on Douglas' ideas. Douglas (2009) suggests that explanatory models and theories can provide scientific understanding of some phenomena thanks to the epistemically crucial functions of predictions that test explanations: they «assist our explanatory endeavors by providing a check on our imagination, helping to narrow the explanatory options to those that will provide a more reliable basis for decision making» (Douglas, 2009: 446). Since according to both De Regt (2017) and Khalifa (2017) understanding is equated to having an explanation, if we take explanation as the epistemological hinge between prediction and understanding, the result is that prediction «enhance our understanding by telling us which of our explanations are the correct ones, and which are not»

(Findl and Suárez, 2021: 5). But a more precise account of ESU through prediction is given by de Regt (2017).

Let's recall de Regt's CTSU described in Chapter 4., and in particular his principles CUP and CIT, that are two criteria associated with a definition of intelligibility: the value that scientists recognize to the set of characteristics of a scientific theory that facilitates its use in making models, and provide explanations (De Regt, 2017: 23). Since this set of characteristics is contextual and relational, he cannot specify the sufficient and necessary conditions that make a theory intelligible. He claims that the intelligibility of a theory «implies that it should be possible to grasp how its predictions are generated» (De Regt, 2017: 102). Considering that scientists are the creators of models, constructors of explanations, and ultimately accountable for predictions, it can be contended that the former statement implies that if a theory is understandable to a scientist, then she can generate predictions from it. This interpretation suggests that possessing intelligible theories is adequate for generating predictions. Let's term this attribute as the prediction-generating character of intelligible theories (Findl and Suárez, 2021: 6). The central question now arises: Is it possible, and if so, how, to transition conceptually from the prediction-generating character of intelligible theories to their capacity to offer explanatory scientific understanding of specific phenomena (referred to as CUP)? De Regt answers affirmatively, in virtue of his assumption holding that there is an «inherent connection between prediction and explanatory understanding» (De Regt, 2017:

107). He later on defends the idea that there is no prediction without understanding: «prediction turns out to be impossible without understanding» (de Regt, 2017: 107). This is a thesis used to defy the possibility of descriptive understanding entailing a prediction without explanatory information concerning a phenomenon: «Perhaps it is possible to devise a purely phenomenological model of a phenomenon, which does not relate to any theories at all, but such a model would merely have a descriptive and perhaps predictive value but yield not explanatory understanding» (de Regt, 2017: 98). He reinforces this idea with the example of an oracle: «An oracle is nothing but a black box that produces seemingly arbitrary predictions. Scientists want more than this: in addition they want insight, and therefore they need to open the black box and consider the workings of the theory that generates the predictions» (de Regt, 2017: 101–2). The example of the oracle is the perfect analogy for what happens with AlphaFold models. We have a predictive model that has a powerful force in generating the structure of the proteins we want to study, yet they do not provide relevant explanatory information about why the proteins fold in such and such way. Examples of descriptive understanding can be found in Findl and Suárez (2021), Galli (2023) and Galli (2024c *forthcoming*).

From the example of the oracle, Findl and Suárez (2021) advance a case study of «early COVID-19 models», that «were what epidemiologists call *statistical models*» (2021: 3). In particular, they focus on the University of Washington's Institute of Health Metrics

and Evaluation (IHME) models in the first stages of COVID-19 infection. Albeit the models were statistical, they had a predictive function, such that «political decision making was informed by estimations derived from *purely predictive* epidemiological models» (*Ivi*). The analysis they give is a useful introduction to the relation between prediction and understanding: «From a philosophical perspective, this form of modelling also raises an interesting question about the relationship between the scientific capacity to predict a phenomenon and the ability to understand it; a topic that had already stimulated the interest of philosophers» (Findl and Suárez, 2021: 3), as De Regt (2017), (Dieguez, 2013), Douglas (2009), Elgin (2017), Frigg and Hartmann (2020), Potochnik (2017), and scientist (Shmueli, 2010). As Findl and Suárez (2021) argue, in early COVID-19 models, prediction and understanding are in an intimate dialectical relation, which is not mediated by explanation, but description. The definition of Descriptive Understanding (DESC) given by (Findl and Suárez, 2021: 20) is the following:

- DESC: A scientific community has descriptive understanding of a phenomenon *P* when they have a model or theory that can generate non-counterfactual predictions of the dynamics that P will follow (i.e., how the values of P will develop over time) and is built on a set of basic empirically-based assumptions A1, A2, A3,, …, that make these predictions plausible.

Their analysis thus far has revealed that the understanding acquired during the construction of the curve-fitting versions of the IHME model lacks explanatory depth. However, a pivotal inquiry remains regarding the nature of understanding derived from the model-building process and its components. They contend that the early iterations of the IHME model, including the April 2020 update, provide descriptive understanding, despite not being counterfactual. Descriptive understanding, in this context, is delineated as follows: when a scientific community possesses a model or theory capable of generating non-counterfactual predictions regarding the dynamics of a phenomenon, based on empirically-grounded assumptions, it signifies a degree of descriptive understanding. Importantly, the degree of descriptive understanding is contingent upon the adequacy of the underlying assumptions, which render the predictions plausible. The subsequent analysis explores the symbiotic relationship between descriptive understanding and the production of predictions within the framework of the IHME model. This relationship is dynamic, with descriptive understanding evolving and enhancing over time.

The genesis of descriptive understanding within the IHME model-building process is elucidated by the integration of a technical framework and fundamental assumptions. The selection of a Gaussian error function as the technical framework stemmed from observations of COVID-19 mortality trends in Wuhan and previous disease outbreaks. This framework, constituting a curve-fitting

approach, was augmented by key assumptions regarding the influence of social distancing measures on mortality rates. These assumptions, combined with the technical framework, facilitated the formulation of a model capable of generating non-trivial predictions regarding mortality rates. These predictions, derived from the integrated model, contribute to the initial stages of descriptive understanding. The iterative refinement of the IHME model underscores the dynamic nature of descriptive understanding. The model's evolution, exemplified by the introduction of a multiple mixture model component to address discrepancies between predictions and empirical data, demonstrates the interplay between assumptions and predictions. By testing predictions against real-world evidence, epidemiologists identified and rectified erroneous assumptions, leading to model improvements. Moreover, the adaptation of the model in response to new insights further enhances its predictive capacity, thereby deepening descriptive understanding.

Critics may question the significance or depth of descriptive understanding within the IHME model, citing its reliance on historical data and simplistic projections. However, descriptive understanding serves as a crucial tool in various scientific domains, facilitating informed decision-making and guiding research endeavors. Moreover, the process of model refinement and comparison with empirical data engenders a nuanced form of understanding that transcends mere prediction generation.

In essence, the IHME model exemplifies a novel modality of understanding, characterized by its descriptive nature and reliance on statistical associations. Despite its departure from traditional explanatory frameworks, descriptive understanding proves indispensable in navigating complex phenomena such as the COVID-19 pandemic. By elucidating the intricate relationship between descriptive understanding, predictions, and model refinement, our analysis underscores the epistemic significance of this emerging paradigm in scientific inquiry.

This account widens the possibility for a descriptive account of SU, which can be useful to foster the analysis of Deep-learning models used in AI research.

9.5. Scientific Understanding and Deep Learning Models

As we have seen, there is a relatively new algorithmic technology that is on the edge of almost every discussion about AI and the future of machine learning, which are deep learning models, or deep neural networks (DNNs). Although the use of DNNs is common for developers, they are not understood well. In particular we di not understand yet what they are and how they work. Still, scientists use DNNs to obtain understanding of some facts, phenomenon, objects.

Between the scholars studying the nature and properties of DNNs, I think Sullivan (2021, 2022, 2023) poses the right question about SU

and ML models to start with: how much detail about the model needs to be known to understand phenomena with ML models? Her answer is that it is largely an external problem of link uncertainty (LU). We have seen how ML techniques, Deep learning in particular, are attracting philosophical attention. I have proposed a way in which ML-trained algorithms, AlphaFold models, can fit in with existing accounts of scientific models and representations for understanding. After the defence of a realist account of DL models, now I submit that also this kind of models can foster SU, in virtue of the representational accuracy of phenomena. The basic idea is that understanding requires a representational link with the target system, and not just a representational link, in fact it may be the wrong way to an inaccurate representation of the phenomena, but also an accurate representational link to the target system.

According to many scholars simple idealized models enable understanding by reducing complexity (Bokulich, 2008; Khalifa, 2017; Potochnik, 2017; Strevens, 2008). In contrast to the idea that also ML models, as idealized models, can reduce complexity, Sullivan (2022) argue that ML models enable understanding differently from the simple idealized models. According to Emily Sullivan, machine learning models do not reduce complexity. Instead, she contends that they merely shift the complexity from one domain to another. While traditional scientific models often simplify complex systems to make them more manageable, machine learning models operate differently. Sullivan suggests that these models don’t

necessarily simplify the underlying complexity of the phenomena they seek to understand or predict. Instead, they encode that complexity into their structure and parameters, often in ways that are difficult for humans to interpret directly.

In Sullivan's view, while machine learning models may appear to provide straightforward predictions or classifications, their inner workings can be opaque and intricate. The complexity inherent in these models lies in the relationships and patterns they learn from vast amounts of data, which may not be readily understandable or explainable in traditional terms. Thus, rather than reducing complexity, machine learning models transform it into a different form that may be challenging for humans to grasp intuitively.

Furthermore, Sullivan highlights that the complexity of machine learning models can introduce new challenges, such as issues of bias, interpretability, and robustness. These models may capture subtle correlations in the data that humans overlook or fail to understand fully, leading to potential ethical or practical implications. Therefore, Sullivan's perspective suggests that while machine learning models offer powerful tools for analysing complex phenomena, they do not necessarily simplify or reduce the underlying complexity but rather reframe it in a different context.

There are mainly two views about understanding with DNNs: Sullivan's and Räz and Beisbart's. Sullivan (2022) argues that scientific understanding with DNNs is not limited by our lack of understanding regarding DNNs. On the contrary, Räz and Beisbart

(2022) claim that understanding DNNs is important and that our current lack of understanding of DNNs actually limit the ability to understand with DNNs.

Emily Sullivan's view contrasts with that of Beisbart and Raz regarding understanding from deep learning models in several key aspects. Sullivan argues that deep learning models do not simplify or reduce the complexity of the phenomena they aim to understand. Instead, they encode this complexity into their structure and parameters, making their inner workings opaque and challenging for humans to interpret directly. She suggests that while these models may provide predictions or classifications, their complexity remains intact, albeit in a transformed form.

On the other hand, Räz and Beisbart's advocate for a more nuanced perspective that takes into account both understanding deep learning models (DLMs) and understanding with DLMs. They propose a pluralistic approach that integrates different conceptions of scientific understanding, including descriptive and explanatory understanding. Unlike Sullivan, Räz and Beisbart emphasize the potential for deep learning models to contribute to scientific understanding, particularly when coupled with human expertise and interpretation. They argue that while DLMs may not offer traditional explanatory understanding in the sense of simplifying complex systems, they can still provide valuable insights and predictions when used in conjunction with other forms of knowledge and expertise.

In summary, Sullivan's view emphasizes the inherent complexity of deep learning models and their limited capacity to simplify complex phenomena, while Beisbart and Raz advocate for a pluralistic approach that recognizes the potential for DLMs to contribute to scientific understanding in conjunction with other forms of knowledge and interpretation

We have seen that scientific models are tools for understanding. ML models are not an exception to that. There are although important differences between ML methods and previous methods of modelling. First of all, as we have seen, ML models are data-driven instead of theory or hypothesis-driven. They are very complex and opaque, due to the black-box property. According to Sullivan (2022) ML complexity and opacity do not get in the way of understanding phenomena so long as the link between the model and the target system does not have a high degree of link uncertainty (LU). Sullivan highlights that «ML models are opaque due to *implementation* opacity (i.e. how ML algorithms and trained models implement functions), and that such opacity is not, in principle, a barrier to understanding phenomena with ML models» (Sullivan, 2023: 341). Sullivan's assertion regarding the opacity of machine learning (ML) models due to their implementation opacity raises important considerations for our understanding of phenomena using these models. While ML algorithms and trained models may operate in opaque ways, Sullivan suggests that this opacity does not inherently hinder our ability to comprehend phenomena with ML

models. This perspective challenges traditional notions of understanding, which often prioritize transparency and interpretability. Instead, Sullivan invites us to embrace the complexity encoded within ML models and explore new avenues for understanding. In doing so, we may discover that while the inner workings of ML models remain opaque, they still have the capacity to provide valuable insights and predictions about complex phenomena. Therefore, rather than viewing opacity as a barrier, we should recognize it as a feature of ML models and seek alternative approaches to harness their predictive power and contribute to our understanding of the world. Ultimately, Sullivan's perspective encourages us to embrace the complexity inherent in ML models and explore innovative ways to leverage their capabilities for scientific inquiry and discovery.

9.6. Understanding with Models: Explanation and Prediction

De Regt's conception of intelligibility goes in the direction of the priority of theoretical constituents over the models' structures. According to his view, it in virtue of the theory being intelligible that scientists can construe models to explain phenomena. In the case we have seen used by De Regt and Khalifa, i.e. the intelligibility of wave mechanics over matrix mechanics, or the Feynman's model over the mathematical intricacies of Bjorken's explanation, it turns out that

the first steps (according to EKS Model) concerns the knowledge of scientific theories, meaning that the theoretical information is necessary to build a model to explain phenomena. In the case of AlphaFold models I discuss in this chapter, the specific aim of the models is not to *prima facie* to explain phenomena, but to predict a specific phenomenon, which is the dynamic folding of a certain protein (or set of proteins, actual or potential, as we have seen). With the theoretical and experimental knowledge (along with the scientific skills researchers of DeepMind team and biologists) the experts have, they are eligible to interpret the structure prediction of proteins, given by AF models, as a possible explanation of the function of the proteins. It means that explicability of a model assumes different forms, according to the skills, background knowledge, and purposes of the scientists working on such models. The basic idea I have defended in this chapter is although that, *pace* the important pragmatic detour offered by De Regt (2017), also ML models of certain kind, i.e. AF models, can be used as a tool to achieve SU, thanks to their representational properties, which are a substantial features of the explanations scientists can derive, describe, advance with the models in play, even if they appear to be predictions (with a form of possible explanations).

According to computational biologists, AF models cannot afford to give use information about specifically how and why proteins fold in a certain way. This means that AF models do not provide insights

about thermodynamic of kinetic features of protein to which we can extrapolate explanations about how and why they fold.

In conclusion, the relationship between prediction and descriptive understanding, as explored through De Regt's conception of intelligibility and exemplified by the case of AlphaFold (AF) models, reveals a nuanced interplay between theoretical knowledge, model construction, and explanatory potential. De Regt emphasizes the primacy of theoretical constituents in enabling scientists to construct models that explain phenomena. This perspective underscores the importance of theoretical and experimental knowledge in shaping the interpretability of models and their capacity to provide explanations. In the case of AF models, while their primary aim may be prediction rather than explicit explanation, their representational properties still allow for the derivation and advancement of explanations by skilled researchers. However, it's crucial to acknowledge that AF models, according to computational biologists, have limitations in providing detailed insights into the thermodynamic and kinetic features of protein folding, which are essential for comprehensive explanations. Thus, the relationship between prediction and descriptive understanding varies depending on factors such as background knowledge, scientific skills, and the specific aims of the researchers. Despite these complexities, the discussion highlights the potential of ML models, like AF, to contribute to scientific understanding through their representational properties, even if their primary function appears to be prediction. This underscores the need for a nuanced

approach that considers the diverse ways in which models can facilitate understanding within the scientific community.

Chapter 10:
Scientific Understanding and Scientific Realism

The aim of this chapter is to foster the study of the interplay between scientific understanding and scientific realism. To do so, I propose to develop some of Alai's ideas and strategies in order to defend a realist approach to scientific understanding. Although the notion of scientific understanding is not central to his work, I think his defence of scientific realism is helpful in disambiguating and defining the role of scientific understanding in arguing for or against scientific realism. In particular, I argue that, in order to clarify their positions, philosophers of understanding should define them with respect to the realism-antirealism debate before digging into the subtleties of understanding (also scientific)[52]. So, the focus here will be on the scientific realism issues, rather than on the Quinean themes which

[52] See, for example, the discussion about scientific understanding and scientific realism in Part II, *Understanding and Scientific Realism* (Lawler, Khalifa and Shech, 2023: 133-214.)

determined my first encounter with Alai[53]. In this chapter, I sketch the main ideas taken from the debate about scientific understanding. Furthermore, I present Alai's contribution to the scientific realism debate, namely his definition of deployment realism (DR), as an upgrading to resist anti-realist criticisms (Lyons 2002, 2016; Laudan, 1981). In the following section, I present the defence of realism provided by Pincock (2023) on the basis of an analysis of scientific understanding. To conclude, I propose a possible line of connection between the debate about scientific realism[54] and scientific understanding and describe the role of essentiality in scientific understanding.

10.1. Scientific Understanding and Scientific Realism: An Opening Connection

In the last twenty years, some scholars, called "friends of understanding" (Khalifa, 2023: 33), discussed the role and the nature of understanding in epistemology, and that of scientific understanding in the philosophy of science. After the long dominance of the notion of explanation, they began to investigate which type and nature is the understanding involved in scientific research. The

[53] See Galli (2024).
[54] For a detailed analysis of the contemporary debate about scientific realism, see Lyons and Vickers (eds.) (2021), and Angelucci, Fano, Ferretti, Galli, Graziani and Tarozzi (2024).

first step towards the connections between SU and SR concerns the basics of both areas: scientific realists defend a thesis about the epistemic success of science and friends of scientific understanding defend a thesis about the aim of science. According to De Regt (2017), understanding is a specific aim of science. According to Alai (2021a; 2023), in line with realism, science succeeds in giving us true, or approximately true, theories: «it would be a miracle if theories were false, yet got right so many novel and risky predictions. Hence, predictively successful theories are true». (Alai, 2021a, p. 183). De Regt describes the success of sciences in terms of understanding, Alai in terms of the truthfulness of the theories. Thus, we see two opposite directions of analysis, one holding the centrality of understanding in science, the other of scientific knowledge. De Regt, indeed, argues against the realist idea that the success of sciences has to do with true theories. On the other side, Alai and the realists argue in favour of the role of truth for the success of science. Also, Potochnik recognizes that the object of scientific knowledge (SK) is not the same as scientific understanding: «Knowledge and understanding go hand in hand, but there is a gap between their objects» (Potochnik, 2020, p. 942). I think it is correct to highlight the difference between SK and SU. This difference is at the heart of the specific strategies, respectively, of realists and friends of understanding. Bridging the gap between the different strategies leads us to the core assumption supporting SR and SU.

Khalifa's (2017, 2023) and De Regt's (2005, 2017) accounts of scientific understanding address and seek to clarify several key issues in the philosophy of science: besides the nature of scientific understanding, the role of models, and the structure of explanations in science. They agree on some important issues but disagree on other equally important ones. For example, they both argue that scientific understanding is not merely a matter of acquiring knowledge or information. Instead, it should be seen as an epistemic achievement involving cognitive skills and epistemic virtues. Understanding is an active process that goes beyond passive knowledge. It is, instead, on the role of SK that they disagree: De Regt argues that understanding has to be decoupled from knowledge, while Khalifa defends the role of SK in SU.

Moreover, both of them emphasize the central role of models in scientific understanding. Models are simplifications of reality that allow scientists to grasp complex phenomena. Understanding is closely tied to how well one can manipulate and make predictions with these models.

Furthermore, Khalifa and De Regt address the challenges posed by scientific explanation. They argue that understanding is not solely dependent on having an explanation in a traditional sense but is also associated with having a good grip on the model and its mechanisms. This challenges the view that understanding is exclusively linked to explanation.

In addition, they acknowledge that SU is often context-dependent and that there can be multiple models and forms of understanding of the same phenomenon. Different models may provide distinct but valid forms of understanding, and scientists often switch between them depending on the problem at hand.

Besides, Khalifa and De Regt stress the heuristic value of models in science. Models are not just tools for explanation but also for exploration. They facilitate scientific discovery and the development of new theories. Also, they propose that scientific understanding often involves integrating different perspectives and models. The ability to switch between models and see how they relate to each other is crucial for a holistic understanding of complex phenomena.

In conclusion, their account challenges, but in opposite directions, the received view of scientific understanding, which traditionally emphasizes a reliance on explanations and conceptual grasp. According to the received view, understanding is a type of knowledge. Lipton and Salmon are the main advocates of this view: «Understanding is not some sort of super-knowledge, but simply more knowledge: knowledge of causes» (Lipton, 2004, p. 30). In addition, Salmon describes the state of the art among philosophers of science focusing on explanation (Salmon, 1989, p. 134-5):

> explanations enhance our understanding of the world. Our understanding is increased (1) when we obtain knowledge of the hidden mechanisms, causal or other, that produce the phenomena we seek to explain, (2) when our

knowledge of the world is so organized that we can comprehend what we know under a smaller number of assumptions than previously, and (3) when we supply missing bits of descriptive knowledge that answer why-questions and remove us from particular sorts of intellectual sorts of intellectual predicaments.

In these quotes both Lipton and Salmon identify understanding with knowledge of an explanation; the former adopts a causal model of explanation, while the latter accepts the ontic/causal, inferential and pragmatic model of explanation. Now, after a few decades, while De Regt argues that the received view is inadequate for understanding the dynamics of scientific practice, Khalifa instead develops his account as a «more regimented descendant of the received view» (Khalifa, 2017, p. 16).

In summary, Khalifa and De Regt's account of SU challenges certain traditional notions of explanations and knowledge and expands the conception of what it means to understand scientific phenomena. They emphasize the role of models, explore the practical and epistemic aspects of understanding, and reassess the importance of explanations in scientific thought.

I'll come now to Alai's claims about scientific realism and his defence strategy.

10.2. Alai's Contribution to the Scientific Realism and Anti-Realism Debate

During his career, Alai contributed to many areas of philosophy: philosophy of language, epistemology, metaphysics, philosophy of AI, and scientific realism. The contributions to this volume cover practically all the issues he discussed in the last years. In my view, Alai's pages on scientific realism are a compelling turning point for everyone interested in that debate. In particular, deployment realism (DR) is a contemporary version of scientific realism that attempts to address some of the challenges posed by anti-realists, particularly concerning the history of science. Beginning with Kitcher and Psillos, deployment realists have been arguing that we can justifiably believe in the truth of the theoretical constituents that are deployed in, or responsible for, the key predictive successes of scientific theories. They do not claim that we can justifiably believe in the truth of scientific theories as wholes or that we can ever be completely certain of the truth of any theoretical constituent. Deployment realists are motivated by the following observations:

a. Scientific theories achieve remarkable predictive successes.

b. These predictive successes are not satisfactorily explained by anti-realist accounts of science.

c. The theoretical constituents that are deployed in, or responsible for, these predictive successes are often retained from

one theory to the next, even when the theories themselves are superseded.

Deployment realists conclude that these theoretical constituents must be approximately true, even if the theories in which they are embedded are not.

Deployment realism is a relatively new position in philosophy of science, still under development. However, it has been gaining traction in recent years, as it offers a promising way to address some of the challenges posed by anti-realists to scientific realism. Here are some of the advantages of deployment realism:

a. It can explain the predictive successes of science without having to commit to the truth of entire scientific theories.
b. It can account for the fact that scientific theories often change over time, while the theoretical constituents that are responsible for their predictive successes are often retained.
c. It is consistent with the history of science, which shows that many successful scientific theories have been superseded by new theories that are more accurate and comprehensive.

However, deployment realism also faces some challenges:

a. It can be difficult to identify which theoretical constituents are responsible for the predictive successes of a scientific theory.

b. Some theoretical constituents may be approximately true without being literally true.

c. Deployment realism involves a careful balance between using theoretical entities for pragmatic purposes and avoiding unwarranted ontological commitments. Philosophers need to explore the criteria for assessing the ontological implications of theoretical terms and the role of empirical success in justifying their use .

d. Another challenge revolves around the metaphysical interpretation of theoretical entities and their epistemic status. What metaphysical status should deployment realists attribute to these entities? And how should they conceive the epistemic justification for the use of theoretical entities, especially when they may be seen as instrumental or heuristic rather than as genuinely representative of the underlying structure of the world?

According to Kitcher (1993) and Psillos (1999), deployment realists are committed only to the «particular hypothesis which was deployed in deriving novel predictions» (Alai, 2021a, p. 184). More precisely, they are committed only to the hypotheses that were used essentially in deriving novel prediction (Alai, 2021a, p. 185): a hypothesis is most probably true only if it was essential in deriving a novel prediction. (In addition, essentiality plays also a (negative) role in defining the novelty of predictions: a predicted phenomenon is novel only if it was not used essentially in building the theory which predicts it).

Psillos (1999) argues that there are two conditions according to which a hypothesis H is deployed essentially in deriving a novel prediction (NP):

1) NP follows from H, together with some other hypotheses Ohs and auxiliary assumptions AA, but not from OHs and AA alone;
2) No other hypothesis H* is available which can do the same job as H, viz. is
a. Compatible with OHs and AA,
b. Non-ad hoc,
c. Potentially explanatory, and
d. Together with OHs and AA predicts NP.

The conditions 1) and 2) define the essentiality of H. However, according to Lyons (2006; 2009), due to its vagueness, this criterion is not applicable to any historical case: it is not clear when H* should be unavailable. Therefore, Lyons' criticism urges deployment realists to discard the essentiality requirement entirely. He concludes that the No Miracle Argument does not work even when restricted to particular (essential) hypotheses, and DR is false (Lyons, 2006, p. 557).

In response to Lyons' objections to DR, Alai works out a more sophisticated proposal to define essentiality. The idea is to substitute the crucial condition 2) in Psillos' criterion with the following:

2') There is no other hypothesis H* which is proper part of H (hence weaker than H) which together with OHs and AA entails NP (Alai, 2021a, p. 188).

Alai claims that this condition captures the Occam's requirement that «H is not redundant, i.e., that one could not explain the derivation of NP by assuming the truth of something less than H» (Alai, 2021a, p. 190). One example proposed is a possible alternative to Newton's gravitational theory:

H. Inside each massive body there resides a demi-god, which attracts the demi-gods dwelling in each other body by a force F = Gm1m2/r2

H is an example of redundant hypothesis, not essential to the predictions of Newton's actual theory; it would predict the same novel phenomena as the Newtonian theory, but there is no need to believe it. In fact, as Alai notes (2021a: 190), H consists of these assumptions:

H*) each body attracts all other bodies by a force F = Gm1m2/r2
Hd-g) F is exerted by a demi-god residing inside each body.

In the example, the hypothesis H would violate condition 2') proposed by Alai «because there would be another hypothesis H*

(i.e., Newton's actual theory) which is a proper part of H and sufficient to derive its novel predictions» (Alai, 2021, p. 190). The alternative condition 2') resists Lyons' objections. Alai then shows how 2') rules out other false assumptions deployed to derive novel predictions, which Lyons cites as counterexamples to DR (Alai, 2021a, p. 191-9).

Although refined by Alai's defence against Lyons, the notion of essentiality was already a key element of Psillos' (1999) position. In fact, despite being often overlooked, this notion is of great interest both for the realists (especially those interested in scientific understanding), and for the friends of understanding. To reconnect the threads of realism and understanding, it is helpful to discuss the Pincock's strategy for a defence of SR based on SU.

10.3. Pincock's Defence of Realism

The debate about the connections between SU and SR has not been settled, yet. A lot of issues are still under examination. To be noticed is the discussion between Pincock (2023) and Potochnik (2023) about the role of scientific realism for SU. Pincock (2023) presents a defence of SR based on the success of science and then defines what features must have SU for this defence to be successful. He advances a minimal definition of what does it mean to be a scientific realist: «someone is a scientific realist just in case they

believe that they know of various unobservable entities and some of their characteristics, and that this knowledge is based on scientific investigations» (Pincock, 2023, p. 135). The example he uses to illustrate the scientific realist position comes from electrostatics. Electrostatic phenomena regard microscopic objects with positive and negative electric charges; these objects cannot be observed. The realist claims that we know the existence and interactions of these charged objects, such as that charges repel each other, and opposite charges attract one another via a force determined by Coulomb's law: F = k (q1q2)/r2. This example triangulates between the practice of the experimenter in scientific inquiry, the knowledge of unobservable entities, which becomes sharable scientific knowledge, and the scientific realism inspired by this typical scientific procedure. According to Pincock,

The correctness of scientific realism requires knowledge of unobservable entities, and so a defense of scientific realism will involve a defense of some knowledge of unobservable entities. This defense takes the form of an argument whose conclusion is that there is knowledge of unobservable entities (by scientific means) (2023, p. 136).

The core of his defence is an inference from the objectual understanding of a phenomenon, i.e. electrostatic induction, to the knowledge of the existence and character of unobserved objects,

such as charged particles. What he calls “objectual understanding” is a demanding kind of cognitive relation between an agent and an object. The inference Pincock builds is

legitimate to the extent that the understanding of a phenomenon involves knowledge of the existence and character of some unobservable entities. A successful defense of scientific realism may then proceed by making it plausible to a neutral party that this sort of understanding exists (2023, p. 136).

The example of electrostatics comes from Feigl’s illustration of “the Copernican turn” (Feigl, 1950, p. 41; Psillos, 2011, p. 308), which requires «a scientist to take whatever evidence they have assembled in favor of some theoretical claims, and to use their theories to explain that evidence and the scientist’s knowledge of that evidence» (Pincock, 2023, pp. 137-8). Fiegl example is the following (Feigl 1950, p. 40):

The divergence of the gold leaves of the electroscope which epistemically serves as an indicator of the presence of charges is immediately deducible from the theoretical assumptions of electrostatics, i.e. primarily from the Coulomb law of attraction and repulsion.

Concerning this deduction, Pincock stresses that its crucial point is the explanation of the evidence together with the scientific knowledge of the evidence (Feigl, 1950, p. 41):

If knowledge (as behaviour) is not to remain an utter mystery or miracle, it is clear that the knowing organism itself must find a place in the world it knows. Whatever object can be reached by empirical knowledge must, no matter how indirectly, be related, (yes, causally related) with the processes in the knowing organism.

Moreover, Pincock notices that Feigl's case of experimental manipulation and epistemic creation of electrostatic induction fits well into Psillos' argument for selective realism. We have already seen an example of DR taken from Alai (2021a) of a specification and correction of the Psillos' strategy in detecting the essentiality of hypotheses. Now we trace the argument for selective realism embedded in Feigl's case (Pincock, 2023, p. 138):

1. The experimenters build their experimental apparatus using the electromagnetic theory, drawing on Coulomb's law for electromagnetic forces.
2. The agent models the metals in this apparatus as conductors, with charged particles that are free to move on the surface of these metals in response to charged bodies being placed near the apparatus.

3. When the gold leaves move as the agent expects using their theory, the agent acquires evidence that Coulomb's law is correct, and that the hypothesized charged particles really do exist, even though they are too small to be observed.

4. So, in such circumstances, the best explanation of the experimental manipulation of the gold leaves is that these theoretical claims are true, and that the mechanisms in question really are operating to produce the experimental effects.

Pincock's argument is an intertwined inference that refers to the same issues addressed by Hacking (1983). The discussion about entity realism brought Hacking to his «if you can spray them, then they are real». He advances a criterion of manipulative success which aims at substituting the explanatory virtue in the justification of scientific belief. Pincock wants to link the manipulative success of the experimenter to the scientific understanding of the phenomenon under scrutiny. But this argument by itself cannot resist various antirealist objections. For instance, an instrumentalist can argue that successful understanding of the phenomena is possible even without the realist claim about the existence of theoretical entities. Thus, working out a framework which allows to connect DR to SU will strengthen the realist position about understanding.

To a closer look, in this example the experimenter has already a theoretical framework to be used to derive from the experiment some new knowledge about a specific phenomenon, viz. electrostatic

induction. There are two types of unobservables in the example: Coulomb's law for electromagnetic forces and the hypothetical charged particles. Without discussing now the role and the different features of the specific law in play and the unobservable charged particles, let us focus on Pincock's argument based on SU and see what role it has in his defence of SR. His general argument is the following (Pincock 2023, pp. 140):

1. An agent conducts an experiment that successfully creates and manipulates an instance of some phenomenon, drawing in part on their theoretical beliefs concerning this phenomenon.
2. The agent grasps that the phenomenon in question obtains independently of their actions and scientific community.
3. The agent grasps that the observable features of this phenomenon depend on the existence and character of some unobservable entities that are posited by their theoretical beliefs.
4. The agent has a good understanding of the phenomenon.
5. Therefore, the agent knows of the existence and character of some unobservable entities.

Another vulnerability in this argument is that understanding and knowledge have been used to define the one in virtue of the other; in fact, we can conceive 1-5 as a definition of knowledge. The circularity of the definition of both understanding and knowledge does not help us to define the features of SU as distinct from

knowledge. In fact, Pincock specifies that the experimenter acquires knowledge of unobservable entities through the understanding of the phenomenon. The understanding, then, arises partially through the awareness of the independence of that phenomenon from the agent and the scientific community: «in the case of electrostatic induction, the experimenter came to know of the existence of unobservable charged particles only because they grasped that the phenomenon of electrostatic induction depends on those particles in an objective way that is valid quite generally» (Pincock, 2023, p. 142). The phenomenon has an independent and objective character, described in part by Coulomb's law and the electromagnetic theory. So, there are three main points to be highlighted in this argument: an independence condition under which the phenomenon obtains, the understanding of the phenomenon and the knowledge about the existence and character of some unobservable entities. The independence condition (IC) on understanding is then defined by Pincock (2023, p. 142):

IC) An agent's understanding of some phenomenon X satisfies the independence condition just in case the agent grasps that X may obtain with the very same characteristics independently of the agent's actions or the operation of their scientific community.

According to Pincock (2023), denying that understanding can meet the IC makes scientific realism untenable. Clearly IC amounts to one

of the most fundamental pillars of SR, i.e., that there are phenomena obtained independently by our presence, existence, and cognition (often called "metaphysical realism"). Scholars, then, divide in two main groups: who denies the possibility for independent phenomena to be known or understood, and who defends the possibility of knowing and understanding them, even while safeguarding their independence. An objection to IC is that certain phenomena produced only experimentally cannot be found in nature, hence IC is not satisfied, even though scientists understand the phenomena under experimentation. This shows that IC should be reformulated to require simply that X may obtain with the very same characteristics independently of the agent's perceiving, believing or knowing that it obtains.

In the above argument is also embedded a specific feature ascribed to SU by Dellsén's Dependency Model of acquiring understanding (DMA), which Pincock uses to constrain SU to elucidate the causal or dependent nexus of the objects in play in the experimental setting:

DMA) S understands a phenomenon, P, if and only if S grasps a sufficiently accurate and comprehensive dependency model of P (or its contextually relevant parts); S's degree of understanding of P is proportional to the accuracy and comprehensiveness of the dependency model (or its contextually relevant parts) (Dellsén, 2020, p. 1268).

DMA has a limitation in that it defines ‘understanding’ through ‘grasping’, which is roughly synonymous. In order rescue (DMA) from circularity, one should define the notion of grasping so to distinguish it from understanding, but it is not immediately obvious how this can be done. Otherwise, I suggest, a viable solution would be simply to substitute ‘grasps’ with ‘builds’. In this way the first part of (DMA) would read: «S understands a phenomenon, P, if and only if S builds a sufficiently accurate and comprehensive dependency model of P (or its contextually relevant parts)». A dependency model of P, here, is a model describing the nexus of relations and objects in play to obtain P.

Now that many ingredients of the debate about SR and SU have been presented, I trace a possible line of connection between scientific realism and scientific understanding.

10.4. Scientific Realism and Scientific Understanding: Crisscrossing Paths.

The claims of deployment realism and the assumptions concerning scientific understanding are connected in several ways. First, deployment realism is a form of scientific realism, which is the view that scientific theories are, at least approximately, true. Deployment realists argue that we can justifiably believe in the truth of the theoretical constituents that are deployed in, or responsible for, the

key predictive successes of scientific theories. On the other hand, as I conceive it, SU is committed to the truth, or at least the approximate truth, of the constituents of understanding, contextually situated in a specific time and research community. This claim clearly supports the assumption that science can provide us with an accurate understanding of the world. SU implies that science aims at, and succeeds in, giving us, also through theories, an understanding of the phenomena. According to DR, as we have seen, the predictive success of scientific theories is based on the approximate truth of deployed theoretical constituents. From these two premises, we can draw the first connection between SU and DR:

SU&DR-1: if science can provide us with an accurate understanding of the world, and if science is predictively successful, understanding implies the predictive success of science.

Assuming both the supporters of SU and DR are right, it follows that science provides understanding and it provides predictive success. Therefore, it is reasonable to suggest that probably this cannot be a mere coincidence. Hence, both groups should try to account for the connection between understanding and success. For instance, is it the case that understanding favours predictive success? Deployment realists might answer yes, arguing that understanding implies truth, and truth provides success. Conversely, does perhaps understanding involve success? Friends of SU might answer yes, because

understanding is more than a passive state of mind, it has to do with a complex interaction between the subject, information and reality, and success may provide a sort of favourable feedback from reality to one's efforts.

The role of DR is then relevant to obtain scientific understanding of a phenomenon, in so far as we want understanding to be successful and DR explains well how theories in science can be successful.

Second, deployment realists acknowledge that scientific theories often change over time. However, they argue that the theoretical constituents that are responsible for the predictive successes of a scientific theory are often retained from one theory to the next. This suggests that these theoretical constituents are approximately true, even if the theories in which they are embedded are not, which entails that there is cumulative progress in science, so that new scientific knowledge builds on previous scientific knowledge. This, in turn, is consistent with the idea that understanding comes with a degree and that the accuracy and comprehensiveness of understanding improve over time.

Finally, deployment realists stress that scientific theories have remarkable predictive success, which suggests that they capture something important about the world, even if they are not strictly true. At the same time, this explains why science is a valuable tool for manipulating and understanding the world, as held by the friends

of understanding: since it has success, it is useful, and since it grasps some truths about the world, it allows to interact with it.

An example of how deployment realism connects with the assumptions about scientific understanding is Newtonian mechanics, as discussed by Alai. Today, in the light of General Relativity, we consider it strictly speaking false. However, it entails certain true assumptions, which can be roughly summarized as «The movement of physical bodies is due to their masses through a mechanism [actually the curvature of space, not gravitation force] which in particular conditions approximately obeys Newton's law» . Since these assumptions provide us with accurate knowledge about the world, not only they yielded the famous predictive successes of the theory, like the existence, mass and position of Neptune, but allow to make accurate predictions about a wide range of phenomena, such as the motion of airplanes and satellites. Scientific understanding is conceived as entailing a broader framework than scientific knowledge, which could be defined as a subset of SU. Then, in the broader picture of Newtonian mechanics, as a case of scientific understanding of specific phenomena, we can now extract false and true assumptions. Someone stresses the false theoretical assumptions and the falsity embedded in scientific models, to decouple SU from knowledge (De Regt 2005, 2017, 2023), someone else focuses instead on the robust relation between SU and scientific knowledge (Khalifa 2017, 2023). I would suggest that we can admit different degrees of SU, in so far as we work out the false or the true

(approximately true) assumptions of the theories. As DR is a viable account to argue for the success of theories, given their approximately true constituents, it is also a viable way to direct us to a better definition of scientific understanding.

To be more specific, the relevance of DR to SU can also be appreciated in connection with Pincock argument in defence of realism and the electrostatics example. First, the claim that the experimenter acquires knowledge of the existence and character of some unobservable entities is implicit in, or at least compatible with DR. Moreover, at step 4 of his argument, Pincock assumes that the agent has a good understanding of the phenomenon only after securing that the IC condition is satisfied. This means that he recognizes that the scientific understanding of a phenomenon requires the scientist to be aware of the independent existence of the phenomenon also in the absence of experimental manipulation.

Looking closer at the example of electrostatic induction, the experimenter starts testing an instance of a certain phenomenon, drawing from the theoretical knowledge of electromagnetism, specifically from Coulomb's law. When the specific behaviour of the gold leaves is obtained, the agent can interpret it on the basis of previous knowledge and embed it in a wider epistemic network, in order to get an understanding of it.

Now, not all the empirical phenomena explained by Coulomb's law were used by him in conceiving and formulating his law. So, we may well suppose that the behaviour of the gold leaves in this experiment,

and perhaps even the phenomenon of electrostatic induction in general, were not used in building Coulomb's law. In this case, they count as novel predictions in the sense of DR. Accordingly, point 3. of Pincock's argument for the defence of realism based on understanding can be reinforced as follow:

3. The agent grasps that the observable features of this phenomenon (OF) are a novel prediction (NP) of a given theoretical framework (H). Hence, she realizes that scientists cannot have conceived H by chance, but because through their research they discovered certain real unobservable entities and some of their properties (roughly those described by H) which are also responsible for the observable features OF.

Therefore, the recognition that a hypothesis was deployed in a novel prediction can well be embedded in the Dependency Model of acquiring understanding (DMA) proposed by Dellsén and used by Pincock to elucidate the causal or dependent nexus of the objects in play in the experimental setting. In this way DR contributes to upgrading the debate on SU by clarifying what is scientific success and how it takes place.

If we take SU as, basically, building a complex dependency model of a phenomenon, stating the precise relations between the elements of this model is one of the fundamental criteria to be satisfied.

As a serious contribution to the characterization of scientific realism, Alai's work provides also some important prerequisites for a better assessment of the connections between the latter and scientific understanding. In my view, it can help evolve this discussion and possibly offer new strategies to the realists. It can also open up new discussions on the complexities, epistemic features, and pervasiveness of science in our lives.

10.5. Khalifa's and De Regt's Stances on Realism

As in the this last and conclusive Chapter we have read, there is a connection between, SU and Scientific Realism (SR), to be detailed. Here I would like to conclude this Chapter with a reflection on Khalifa's and De Regt's stances on realism, derived from the position they defend as friend and frenemy of the Explanatory model of SU. Khalifa defends what he calls voluntarism, that is a kind of quietism about the explanatory truth-requirement (Khalifa, 2023: 43). Indeed, he submit, as a scholar of the received view, that knowledge and understanding are intimately interwoven: «The least controversial aspect of knowledge is that it is *factive*; i.e. that knowledge that *p* requires *p* to be true. By contrast, understanding and truth do not always cooperate. The march of science is littered with false theories that nevertheless advanced our understanding» (Khalifa, 2017: 154). But on the other hand, de Regt (2015) argues that understanding has

to be decoupled from knowledge and truth, so his view of SU is one of the deflationary and pragmatist alternative to realism.

Khalifa and De Regt offer contrasting perspectives on the relationship between scientific understanding (SU) and scientific realism (SR), particularly regarding the truth-requirements of explanatory models. Khalifa's stance, characterized as voluntarism, suggests a quietism about the explanatory truth-requirement, acknowledging the intertwining of knowledge and understanding while recognizing that understanding does not always necessitate truth. Khalifa acknowledges the factive nature of knowledge but proposes a mediation regarding the truth-requirements of explanations. In contrast, De Regt leans towards a more realist conception of explanation truth-requirements. While Khalifa's perspective emphasizes the complexity of the relationship between understanding and truth in advancing scientific knowledge, De Regt's approach aligns with a more stringent realism, suggesting that explanatory models should adhere closely to truth-requirements. These differing perspectives shed light on the nuanced discussions surrounding the epistemic commitments inherent in scientific understanding and its relationship to realism. Then, Khalifa advances a proposal on a mediation: voluntarism about explanation truth-requirements. In arguing in favour of the deployment realism tenets concerning the requirements on SU, I am inclined to hold a more realist conception of the explanations truth-requirements.

The debate about the realist and antirealist assumptions in the accounts of SU we have seen is still ongoing, and the relation between SR and SU is one of a foundational importance, that is now an open issue, that deserves further studies.

Conclusion

This thesis represents not just a journey through the landscape of scientific understanding (SU) but a call to embrace the richness of its diversity and the urgence to dedicate further studies to understanding and scientific understanding, in both philosophy of science and epistemology. From the foundational inquiries into scientific explanations in Chapter 1 to the nuanced exploration of *Verstehen* in Chapter 2, each step of this journey has illuminated the multifaceted nature of understanding within scientific inquiry. Yet, as we stand at the threshold of a new era marked by rapid technological advancement, it becomes increasingly apparent that our understanding of SU is far from complete.

Chapters 3 and 4 critically engaged with the perspectives of Schurz, Lambert, and De Regt, challenging the confines of the Received View and advocating for alternative frameworks that capture the dynamic interplay between explanation and understanding. Chapter 5 further expanded the discourse by examining Khalifa's insights, shedding light on the intricate relationship between scientific explanation and knowledge. However,

as we delve deeper into the implications of emerging technologies such as AlphaFold models and language models (Chapters 7 and 8), new questions arise about the nature of understanding in the age of Artificial Intelligence.

As we navigate these uncharted waters, it becomes imperative to confront the open issues that linger at the intersection of philosophy of science and Artificial Intelligence. How do we reconcile the predictive power of machine learning models with our traditional notions of scientific understanding? What role do these models play in shaping our epistemic landscape, and how do they influence our perception of explanatory adequacy? These are just some of the pressing questions that warrant further exploration in future studies.

In synthesizing the diverse perspectives and theoretical frameworks presented in this thesis, we are reminded of the profound complexity of SU and the ongoing quest to unravel its characteristics. By advocating for a pluralistic approach — one that embraces both descriptive and explanatory understanding — we lay the groundwork for a more inclusive and robust understanding of scientific understanding. As we embark on the next phase of our intellectual exploration, let us heed the call to embrace the plurality of perspectives and methodologies that enrich our appreciation of the natural and artificial world and pave the way for new frontiers in scientific and philosophical inquiry.

References

Abbott, B. P., *et al*., (2016), "Observation of Gravitational Waves from a Binary Black Hole Merger", *PRL*, 116, 061102, 1–16.

Alai, M., (1992), *Modi del realismo. Soggettività, convenzioni e sostenibilità del realismo*, Milano, FrancoAngeli.

Alai, M., (1998), *Filosofia della Scienza del Novecento*, Roma: Armando Editore.

Alai, M., (2021a), *La Filosofia Analitica del Linguaggio*, Milano: Mimesis.

Alai, M., (2021b), "Scientific Realism and Further Underdetermination Challenges", *Axiomathes*, 31, 779–89.

Alai, M., (2021c), "The Historical Challenge to Realism and Essential Deployment". In Lyons, T.D. and Vickers, P. (eds.), *Contemporary Scientific Realism: The Challenge from the History of Science*. Oxford: Oxford University Press, 183–215.

Alai, M., (2023), "Scientific Realism, Metaphysical Antirealism and the No Miracle Arguments", *Found Sci*, 28, 377–400.

Angelucci, A., Fano, V., Ferretti, G., Galli, G., Graziani, P., Tarozzi, G. (eds.), (2024), *Realism and Antirealism in Metaphysics, Science and Language*. *Festschrift for Mario Alai*. Milano: FrancoAngeli

Ankeny, R., A. and Leonelli, S., (2020), *Model Organisms*, Cambridge: Cambridge University Press.

Annas, J., (1981), *An Introduction to Plato's Republic*, New York: Oxford University Press.

Baker, Gordon P. and Hacker, Peter M. S., (1984), "On Misunderstanding Wittgenstein: Kripke's Private Language Argument", Essays on Wittgenstein's Later Philosophy, *Synthese*, 58(3), 407–50.

Bai, X.C., McMullan, G., and Scheres, S.H.W., (2015), "How Cryo-EM Is Revolutionizing Structural Biology". *Trends Biochem. Sci.*, 40, 49–57.

Bechtel, W. and Richardson, R. C., (1993), *Discovering complexity: Decomposition and Localization as Strategies in Scientific Research*, Princeton: Princeton University Press.

Bengio, Y., Léornad, N, and Courville A., (2013), "Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation", arXiv:1308.3432, (last view, 3/9/2023).

Bianchini, F., (2018), "The Problem of Prediction in Artificial Intelligence and Synthetic Biology", *Complex Systems*, 27(3), 249–265

Boyer-Kassem, T. (2014). "Layers of Models in Computer Simulations". *International Studies in the Philosophy of Science*, 28, 4, 417–36.

Boniolo, G., (1999), *Metodo e rappresentazioni del mondo*, Milano: Bruno Mondadori.

Brézillon, P., Turner, R. and Penco, C. (eds.), (2007), *Modeling and Using Context*. 10th International and Interdisciplinary Conference, CONTEXT 2017, Paris, France, June 20-23, 2017, Proceedings, LNAI, Springer.

Broeks, D., Knuuttila, T., and De Regt, H., W., (2023), "Understanding, virtually: How does the synthetic cell matter?". *Perspectives on Science*, 1–39.

Bromberger, S., (1962), *An Approach to Explanation*. In R.S. Butler (ed.), *Analytical Philosophy, 2nd series*, Oxford: Basil Blackwell.

Brożec, B., Furman, M., Jakubiec, M. and Kucharzyk, B., (2023), "The Black Box Problem Revisited. Real and Imaginary Challenges for Automated Legal Decision Making", *Artificial Intelligence and Law*, 1–14.

Buckner, C. and Garson, J., (2019), "Connectionism", *The Stanford Encyclopedia of Philosophy*, Edward N., Z. (ed.): https://plato.stanford.edu/entries/connectionism/ (last view, 12/09/2023).

Bullmore, E. and Sporns, O., (2012), "The economy of brain network organization", *Nat. Rev. Neurosci*., 13, 336–349.

Campbell, N., R., (1920), *Physics: The Elements*, 1st edition. Cambridge: Cambridge University Press.

Candlish, S., (2004), "Private Objects and Experimental Psychology", in *Wittgenstein Today*, Coliva, A. and Picardi, E., (eds.), 297–317, Padova: Poligrafo.

Carter, J., A. and Gordon, C. E., (2014), "Objectual understanding and the value problem" *American philosophical quarterly* 51 (1), 1–13.

Cartwright, N., (1983), *How the laws of physics lie*, New York: Oxford University Press.

Cartwright, N., Hardie, J., Montuschi, E., Soleiman, M., and Thresher, A.C., (2022), *The Tangle of Science. Reliability Beyond Method, Rigour and Objectivity*, Oxford: Oxford University Press.

Church, K., (2007), "A Pendulum Swung too Far", in *Linguistic Issues in Language Technology*, 2 (4): 1–26.

Cobb, M., (2015), *Life's greatest secret. The race to crack the genetic code*, New York: Basic Books.

Cobb, M., (2020), *The idea of the brain*. *The past and future of neuroscience*, New York: Basic Books.

Contessa, G., (2010). "Editorial Introduction to Special Issue". *Synthese*, 172, 2, 193–95.

Cornelissen, M., D., and De Regt, H., W., (2022) "Understanding in synthetic chemistry: the case of periplanone B", *Synthese*, 200, 461, 1–31.

Craver, C. F., (2005), "Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience", *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 2, 373–395.

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

Cummins, R. C., (1975), "Functional analysis", *J. Philos*., 72, 741–765.

Cummins, R. C., (1983), *The Nature of Psychological Explanation*, Cambridge, MA: MIT Press.

Cummins, R. C., (2000), "How does it work? versus 'what are the laws?': Two conceptions of psychological explanation", in *Explanation and Cognition*, Keil, F. C., and Wilson, R. A. (eds.), The Cambridge, MA: MIT Press, 117–144.

Danielsson, U., (2023), *The World Itself: Consciousness and Everything of Physics*, New York: Bellevue Literary Press.

De Regt, H., W., (1997), "Erwin Schrödinger, *Anschaulichkeit*, and Quantum Theory", *Studies in History and Philosophy of Modern Physics*, 28, 461–82.

De Regt, H., W., and Dieks, D., (2005), "A Contextual Approach to Scientific Understanding". *Synthese*, 144, 137–170.

De Regt, H., W., (2009), “The Epistemic Value of Understanding”, *Philosophy of Science*, 76(5), 585–97.

De Regt, H., W., (2015), “Scientific Understanding: Truth or Dare?”, *Synthese*, 192, 3781–97.

De Regt, H., W., (2017), *Understanding Scientific Understanding*. Oxford: Oxford University Press.

De Regt, H., Leonelli, S. and Eigner, K. (eds.), (2009), *Scientific Understanding: Philosophical Perspectives*, Pittsburgh: University of Pittsburgh Press.

Dellsén, F., (2020), “Beyond Explanation: Understanding as Dependency Modelling”, *British Journal for the Philosophy of Science*, 71, 1261–86.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., (2019), “BERT: Pre-training of deep bidirectional transformers for language understanding”, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT), Association for Computational Linguistics: 1–16.

Diamond, C., (2000), “Does Bismarck Have a Beetle in His Box”, in *The New Wittgenstein*, Crary, A. and Read, R., (eds.), 262–92. London: Routledge.

Dieguez, A., (2013), “When do models provide genuine understanding, and why does it matter?”, *History and Philosophy of the Life Sciences*, 35, 599–620.

Dijksterhuis, E., J., (1950), *De Mechanisering van het Wereldbeeld*, Amsterdam: Meulenhoff.

Dill, K.A., Ozkan, S.B., Shell, M.S., and Weikl, T.R., (2008), “The Protein Folding Problem”. *Annu Rev Biophys.*, 37, 289–316.

Douglas, H. E., (2009), “Reintroducing Prediction to Explanation”, *Philosophy of Science*, 76 (4), 444–63.

Dowe, P., (2000), *Physical Causation*, Cambridge: Cambridge University Press.

Downes, S., (2009), “Models, Pictures, and Unified Accounts of Representation: Lesson from Aesthetics for Philosophy of Science”. *Perspective on Science*, 17, 4, 417–28.

Dror, I. and Dascal, M., (1997), “Can Wittgenstein help free the mind from rules?”, in Johnson, D., Erneling, C. (eds.), *The philosophical foundations of connectionism*, 217–26, Oxford: Oxford University Press.

Ducheyne, S., (2008), “Towards an Ontology of Scientific Models”. *Metaphysica*, 9, 1, 119–27.

Durán, J.M., (2018), *Computer Simulations in Science and Engineering: Concepts—Practices—Perspectives*, Berlin: Springer.

Durán, J.M., (2020), “What Is a Simulation Model?”, *Minds & Machines*, 30, 301–23.

Elgin, C. Z., (1991), “Understanding: Art and Science”, *Midwest Studies in Philosophy*, 16: 196–208.

Elgin, C. Z., (1996), *Considered Judgment*, Princeton: Princeton University Press.

Elgin, C. Z., (2004), “True Enough”, *Philosophical Issues*, 14: 113–121.

Elgin, C. Z., (2007), “Understanding and the Facts”, *Philosophical Studies*, 132, 33–42.

Elgin, C. Z., (2009), “Is Understanding Factive?”, in *Epistemic Value*, 322–330 (Appendix C). Haddock, A., Millar, A. and Pritchard, D., (eds.), New York: Oxford University Press.

Elgin, C. Z., (2010), “Review of De Regt, Leonelli, and Eigner, eds., *Scientific Understanding: Philosophical Perspectives*, *Notre Dame Philosophy Reviews*: https://ndpr.nd.edu/reviews/scientific-understanding-philosophical-perspectives/ (last view 13/4/2024).

Elgin, C. Z., (2017), True Enough, Cambridge, MA: MIT Press.

Elman, J,, L., (2014), “Systematicity in the lexicon: On Having Your Cake and Eating It Too”, in *The Architecture of Cognition. Rethinking Fodor and Pylyshyn’s Systematicity Challenge*, Calvo, P. and Symons, J., 115–46, Massachusetts: The MIT Press.

Elowitz, M.B. and Lim, W.A., (2010), “Build Life to Understand It”. *Nature*, 468, 7326, 889–90.

Facchini, A. and Termine, A., (2022), “Towards a Taxonomy for the Opacity of AI Systems”, in *Philosophy and Theory of Artificial Intelligence 2021. PTAI 2021. Studies in Applied Philosophy, Epistemology and Rational Ethics*, Müller, V.C. (ed.), vol 63, 73–89, Springer, Cham.

Fano, V., (2005), *Comprendere la scienza*, Napoli: Liguori.

Feigl, H., (1950), “Existential Hypotheses: Realistic versus Phenomenalistic Interpretations”, *Philosophy of Science*, 17, 35–62.

Findl, J., and Suárez, J., (2021), “Descriptive understanding and prediction in COVID-19 modelling”, *HPLS*, 43, 107, 1–31.

Firth, J., R., (1968), *Selected papers 1952–1959*, London and Harlow: Longmans, Green and Co Ltd.

Firth, J., R., (1969), *Papers in linguistics 1934–1951*, London: Oxford University Press.

Fischer, E., (1894), “Einfluss der Konfiguration auf die Wirkung der Enzyme”. *Berichte der deutschen Chemischen Gesellschaft*, 27, 3, 2985-93.

Fischer, E., (1906), *Untersuchungen über Aminosäuren, Polypeptide und Proteine*, Vol. 2. Berlin: Springer.

Firedman, M., (1974), “Explanation and Scientific Understanding”, *Journal of Philosophy*, 71, 15–19.

Fodor, J., A., (1968), *Psychological Explanation: An Introduction to the Philosophy Of Psychology*, Random House, New York.

Fodor, J., A. and Pylyshyn, Z., W. (1988), “Connectionism and Cognitive Architecture: A Critical Analysis”, *Cognition*, 28(1–2), 3–71.

Fogelin, R., J. (1976), *Wittgenstein*, London and Boston: Routledge.

Frigg, R. and Nguyen, J., (2017), “Models and Representation”. In Magnani, L. and Bertolotti, T. (eds.), *Springer Handbook of Model-Based Science*. Dordrecht: Springer, 49–102.

Frigg, R. and Hartmann, S., (2020), “Models in Science”, *The Stanford Encyclopedia of Philosophy*, Zalta, E., N. (ed.): https://plato.stanford.edu/entries/models-science/ (last view, 4/5/2023).

Frigg, R. and Nguyen, J. (2022), “Scientific Representation”, *The Stanford Encyclopedia of Philosophy*. Zalta, E., N. (ed.): https://plato.stanford.edu/entries/scientific-representation/ (last view, 9/12/2023).

Galilei, G., [1623] (1960), “Two kinds of properties. Selection from Il Saggiatore”, in *Philosophy of Science*, Danto, A., C., and Morgenbesser, S., (eds.), New York: Meridian Books.

Galli, G., (2023), “Structure Representation of Deep-Learning Models: The Case of AlphaFold”, *Argumenta*, 9, 1, 43–60.

Galli, G., (2024a), “Scientific Realism and Scientific Understanding”, in Angelucci, A., Fano, V., Ferretti, G., Galli, G., Graziani, P., Tarozzi, G. (eds.), *Realism and Antirealism in Metaphysics, Science and Language*. *Festschrift for Mario Alai*. Milano: FrancoAngeli, 155–170.

Galli, G., (2024b *forthcoming*), "Scientific Understanding and the Explanatory Integration in Cognitive Sciences", *Software Engineering and Formal Methods. SEFM 2022 Collocated Workshops*. *AI4EA, F-IDE, CoSim-CPS, CIFMA, Eindhoven, Nederland, November 6–10, 2023, Revised Selected Papers*.

Galli, G., (2024c *forthcoming*), "Language Models and the Private Language Argument: A Wittgensteinian Guide to Machine Learning", in *Wittgenstein and Artificial Intelligence, Volume 1: Mind and Language*, Helliwell, A., Rossi, A. and Bell, B., (eds.), London: Anthem Press.

Gayler, R., W., (2003), "Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience", in *ICCS/ASCS international conference on cognitive science*, edited by Slezak, Peter, 133–38. Sydney, Australia: University of New South Wales, CogPrints.

Gelfert, A., (2016), *How to Do Science with Models: A Philosophical Primer*. Berlin: Springer.

Gentner, D., (2019), "Cognitive science is and should be pluralistic", *Topics in Cognitive Science*, 11(4), 884–91.

Giere, R., (1988), *Explaining Science*. Chicago: University of Chicago Press.

Giere, R., (1999), "Using Models to Represent Reality". In Magnani, L., Nersessian, N.J., and Thagard, P. (eds), *Model-Based Reasoning in Scientific Discovery*. New York: Plenum Publishers, 41–57.

Giunti, M., Garavaglia, F. G., Giuntini, R., Pinna, S. and Sergioli, G., (2023), "Chatgpt Prospective Student at Medical School", (March 5, 2023), available at SSRN: https://ssrn.com/abstract=4378743 or http://dx.doi.org/10.2139/ssrn.4378743

Godfrey-Smith, P., (2006), "The Strategy of Model-Based Science". *Biology and Philosophy*, 21, 725–40.

Goldstein, L. and Slater, H., (1998), "Wittgenstein, semantics and connectionism", *Philosophical Investigations* 21(4), 293–314.

Greco, J., (2010), *Achieving Knowledge. A Virtue Theoretic Account of Epistemic Normativity*, Cambridge: Cambridge University Press.

Grimm, S., (2021), "Understanding", *The Stanford Encyclopedia of Philosophy*, Edward N., Z. (ed.): https://plato.stanford.edu/archives/sum2021/entries/understanding/ (last view, 10/6/2024).

Hacker, P. M. S., (2001), *Wittgenstein: Connections and Controversies*, Oxford: Clarendon Press.

Hacking, I., (1983), *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*, Cambridge: Cambridge University Press.

Hacking, I., (2001), *Façonner les gens*, lesson at Collège de France, https://www.college-de-france.fr/sites/default/files/documents/ian-hacking/UPL7997567846150782232_Hacking2001_2002.pdf (last view, 3/2/2024).

Hempel, C., G., and Oppenheim, P., (1948), "Studies in the Logic of Explanation", *Philosophy of Science*, 15, 2, 135–75.

Hesse, M. B., (1966), *Models and Analogies in Science*, Notre Dame: University of Notre Dame Press.

Hills, A., (2009), "Moral testimony and moral epistemology", *Ethics*, 120(1), 94–127.

Hills, A., (2016), "Understanding Why", *Noûs*, 50(4), 661–88.

Hollister, D. L., Gonzalez, A. and Hollister, J., (2017), "Contextual Reasoning in Human Cognition and the Implications for Artificial Intelligence Systems", in *Modeling and Using Context*, Brézillon, P., Turner, R., and Penco, C., (eds.), 599–608, Cham: Springer.

Hughes, R.I.G., (1997), “Models and Representation”. *Philosophy of Science*, 64, 325–36.

Humphreys, P., (1989), *The Chances of Explanation: Causal Explanation in the Social, Medical and Physical Sciences*, Princeton: Princeton University Press.

Jaskolski, M., Dauter, Z., and Wlodawer, A., (2014), “A Brief History of Macromolecular Crystallography, Illustrated by a Family Tree and Its Nobel Fruits”, *FEBS J.*, 281, 3985–4009.

Jumper, J., Evans, R., Pritzel, A. et al., (2021a), “Highly Accurate Protein Structure Prediction with AlphaFold”, *Nature*, 596, 583–89.

Jumper, J., Evans, R., Pritzel, A. et al., (2021b), “Supplementary Information for Highly Accurate Protein Structure Prediction with AlphaFold”, *Nature*, 596, 583–89.

Kanerva, P., (1994), “The Spatter Code for Encoding Concepts at Many Levels”, *International Conference on Artificial Neural Networks*, ICANN ’94, Springer, 226–29.

Kaplan, D. M., (2011), “Explanation and description in computational neuroscience”, *Synthese*, 183, 339–73.

Kaplan, D. M., (2017), *Explanation and Integration in Mind and Brain Science*, 1st Edn, Oxford, Oxford University Press.

Kargon, R. and Achinstein, P., (eds.), (1987), *Kelvin’s Baltimore Lectures and Modern Theoretical Physics*, Cambridge MA: MIT Press.

Karplus, M. and McCammon J.A., (1986), “The Dynamics of Proteins”. *Scientific American*, 254, 42–51.

Khalifa, K., (2012), “Inaugurating understanding or repackaging explanation?”, *Philos. Sci*. 79, 15–37.

Khalifa, K., (2013a), “Is understanding explanatory or objectual?”, *Synthese* 190, 1153–1171.

Khalifa, K., (2013b), "The role of explanation in understanding", *Br. J. Philos. Sci.*, 64, 161–187.

Khalifa, K., (2016), "Must Understanding be Coherent?", in *Explaining Understanding*, Grimm, S., Baumberger, C., Ammon, S. (eds.), 139–165.

Khalifa, K., (2017), *Understanding, Explanation, and Scientific Knowledge*, Cambridge: Cambridge University Press.

Khalifa, K., (2019), "Is Verstehen scientific understanding?" *Philos. Soc. Sci.* 49, 282–306.

Khalifa, K., Islam, F., Gamboa, J. P., Wilkenfeld, D. A., Kostic, D., (2022), "Integrating Philosophy of Understanding With The Cognitive Sciences", *Frontiers in Systems Neuroscience*, 16, 1–17.

Khalifa, K., (2023), "Should friends and frenemies of understanding be friends? discussing de Regt," in *Scientific Understanding and Representation: Modeling in the Physical Sciences*, Khalifa, K., Lawler, I., and Shech, E., (eds), Routledge: London.

Khalifa, K., Lawler, I., Shech, E., (2023), *Scientific Understanding and Representation: Modeling in the Physical Sciences*, Routledge, London.

Kitaev, N., Cao, S. and Klein, D., (2019), "Multilingual constituency parsing with self-attention and pre-training", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3499–3505.

Kitcher, P., (1993), *The Advancement of Science*, New York: Oxford University Press.

Knanvig, J., (2003), *The Value of Knowledge and the Pursuit of Understanding*, Cambridge, Cambridge University Press.

Knuuttila, T., (2005), “How Do Models Give Us Knowledge? The Case of Carnot’s Ideal Heat Engine”. *European Journal for Philosophy of Science*, 1, 309–34.

Knuuttila, T., (2011), *Modelling and Representing: An Artefactual Approach to Models*. Studies in History and Philosophy of Science. Springer.

Knuuttila, T., (2018), “Models as Tools in Scientific Practice”. In Ratti, E. and Forber, P. (eds.), *Causality in the Sciences*. Oxford: Oxford University Press, 193–212.

Knuuttila, T., (2021), “Epistemic Artifacts and the Modal Dimension of Modeling”, *Euro Jnl Phil Sci*, 11, 65.

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes*. Cambridge: Cambridge University Press.

Laplane, L., Mantovani, P., Adolphs, R., Chang, H., Mantovani, A., McFall-Ngai, M., Rovelli, C., Sober, E. and Pradeu, T., (2019), “Why Science Needs Philosophy”, *PNAS opinion*, 116 (10), 3948–52.

Latora, V., Marchiori, M., (2001), “Efficient behaviour of small-world networks”, *Phys. Rev. Lett.* 87, 198701.

Laudan, L., (1981), “A Confutation of Convergent Realism”, *Philosophy of Science,* 48, 19–49.

Laurents, DV., (2022), “AlphaFold 2 and NMR Spectroscopy: Partners to Understand Protein Structure, Dynamics and Function”, *Front Mol Biosci*. 1–l2.

Lawler, I., Khalifa, K., and Shech, E. (eds.), (2023), *Scientific Understanding and Representation. Modeling in the Physical Sciences*. New York: Routledge.

Le Bihan, S., (2001), “The Many Faces of Renormalization Group Explanation”. *Philosophy of Science*, 68, 3, 531-48.

Lear, J., (1988), *Aristotle: The Desire to Understand*, Cambridge: Cambridge University Press.

Legrenzi, P., (2002), *Prima lezione di scienze cognitive*, Roma: Laterza.

Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., and Pitts, W.H. (1959). "What the Frog's Eye Tells the Frog's Brain". *Proceedings of the Institute of Radio Engineers*, 47, 1940–51.

Lipton, P. (2004). *Inference to the Best Explanation*. London: Routledge.

Lombardini, G. (2019). "Carnot's Heat Engine as a Transcendental Instrument: Artefacts and the Claims of the Artefactual Account". *Philosophy of Science*, 86, 5, 1353–66.

Lowney, C. W., Levy, D., Meroney, Ross W. and G., 2020, "Connecting Twenty-First Century Connectionism and Wittgenstein", *Philosophia* 48, 643–71.

Lyons, T. D., (2002) "The Pessimistic Meta-Modus Tollens", in Clarke, S. and Lyons, T. D. (eds.), *Recent Themes in the Philosophy Science*, *Australasian Studies in History and Philosophy of Science*, vol 17, Dordrecht: Springer, 63–90.

Lyons, T. D., (2016), "Structural Realism versus Deployment Realism: A Comparative Evaluation", *Studies in History and Philosophy of Science*, Part A, 59, 95–105.

Lyons, T. and Vickers, P., (eds.), (2021), *Contemporary Scientific Realism. The Challenge from the History of Science*, Oxford: Oxford University Press.

Machamer, P., Darden, L., and Craver, C.F. (2000). "Thinking about Mechanisms". *Philosophy of Science*, 67, 1, 1–25.

Magnani, L. (2001). *Abductive Cognition: The Epistemological and Eco-Cognitive Dimensions of Hypothetical Reasoning*. Berlin: Springer.

Magnani, L. and Nersessian, N.J. (2002). *Model-Based Reasoning: Science, Technology, Values*. Berlin: Springer.

Marconi, D., (1997), *Lexical Competence*, Cambridge (Mass): The MIT Press.

Marconi, D., (2001), *Filosofia e scienza cognitiva*, Roma: Laterza.

Mikolov, T., Chen, K., Corrado, G., and Dean, J., (2013), "Efficient Estimation of Word Representations in Vector Space": https://arxiv.org/abs/1301.3781.

Mills, S., (1993), "Wittgenstein and connectionism: A significant complementarity?", *Royal Institute of Philosophy Supplement*, 34, 137–57.

Miłkowski, M., (2013), *Explaining the Computational Mind*. Cambridge, Massachusetts: MIT Press.

Miłkowski, M., (2016), "Unification strategies in cognitive science. Studies in Logic", *Grammar and Rhetoric*, 48 (1), 13–33.

Minsky, M., and Papert, S., (1969), *Perceptrons: An Introduction to Computational Geometry*, Massachusetts: The MIT Press.

Mitchell, M., (2019), *Artificial Intelligence. A Guide for Thinking Humans*, New York: Farrar, Strauss and Giroux.

Morgan, M. and Morrison, M. (eds.), (1999), *Models as Mediators*, Cambridge: Cambridge University Press.

Morrison, M. (1999). *Models as Autonomous Agents*. In Magnani, L., Nersessian, N.J., and Thagard, P. (eds.), *Model-Based Reasoning in Scientific Discovery*. New York: Plenum Publishers, 1–23.

Morrison, M. (2015). "When Scientists Choose Models, Will Generality Win?". *Erkenntnis*, 80, 3, 521–42.

Morrison, M. (2020). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Nagel, E. (1961). *The Structure of Science. Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace and World, Inc.

Nersessian, N.J. (2008). *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

Nersessian, N.J. (2012). “The Cognitive Representation of Scientific Practice”. In Dunér, D. and Andler, D. (eds.), *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*. Berlin: Springer, 161–81.

Nersessian, N.J. (2017). “Model-Based Reasoning in Scientific Practice”. In Magnani, L. and Bertolotti, T. (eds.), *Springer Handbook of Model-Based Science*. Dordrecht: Springer, 3–29.

Neurath, O., (1937), Unified science and its encyclopaedia. Philosophy of Science, 4, 2, 265–277.

Newman, M., (2012), “An inferential model of scientific understanding”, *International studies in the philosophy of science* 26(1), 1–26.

Newman, M., (2014), . “EMU and inference: what the explanatory model of scientific understanding ignores”, *European journal for philosophy of science* 4(1), 55–74.

Oppenheim, P., Putnam, H., (1958), *Unity of Science as a Working Hypothesis*, Minnesota Studies in the Philosophy of Science, 2, 3–36.

Park, H.S. (2019). “Top 10 Contributions of Cryo-EM to Structural Biology”. *Nature Communications*, 10, 1–10.

Pavlus, J., (2024), “Does AI Know What an Apple Is? She Aims to Find Out”, interview with Ellie Pavlick, *Quantamagazine*, 25/4/2024: https://www.quantamagazine.org/does-ai-know-what-an-apple-is-she-aims-to-find-out-20240425/?mc_cid=f989822778&mc_eid=370b9273ba (last view 25/4/2024).

Perconti, P. and Plebe, A., (2020), “Deep learning and cognitive science”, *Cognition*, 203, 104365,.

Pessoa, F., [1982] (2015), *The Book of Disquiet*, London: Penguin Classics.

Piccinini, G. and Craver, C., (2011), "Integrating psychology and neuroscience: functional analyses as mechanism sketches", *Synthese,* 183, 283–311.

Picollet-D'hahan N., Dolega, M. E., Freida, D., Martin, D. K., Gidrol, X., (2017), "Deciphering cell intrinsic properties: A key issue for robust organoid production", *Trends Biotechnol I*, 35, 1035–48.

Pincock, C., (2023), "Understanding the Success of Science", in Lawler, I., Khalifa, K. and Shech, E. (eds.), *Scientific Understanding and Representation. Modeling in the Physical Sciences*, London-New York: Routledge, 135–150.

Pitt, J.C., (2004), "The Ontology of Mechanisms", *Philosophy of Science*, 71, 1, 1–25.

Pitt, J.C., (2009), "Understanding and Explanation", in D. Dieks, P. Bokulich, S. Hartmann, T. Uebel, M. Weber (eds.), *The Present Situation in the Philosophy of Science*, Dordrecht: Springer.

Plate, T., (2003), *Holographic reduced representations*, Stanford: CSLI Publications.

Poe, E., A., (1968), "On the impossibility of writing a truthful autobiography", in *The Portable E. A. Poe*, New York: Viking Portable.

Potochnik, A., (2017), *Idealization and the aims of science*, Chicago: University of Chicago Press.

Poveda, J. and Vellido, A., (2006), "Neural Network Models for Language Acquisition: A Brief Survey", in *IDEAL Intelligent Data Engineering and Automated Learning*, Corchado, E., Yin, H., Botti, V. and Fyfe, C. (eds.), Berlin: Springer.

Pritchard, D., (2005), *Epistemic luck*, Oxford: Oxford University Press.

Pritchard, D., (2008), "Knowing the answer, understanding, and epistemic value", *Grazer Philosophische Studien*, 77, 325–339.

Pritchard, D., (2009), “Safety-based epistemology: whither now?” *Journal of philosophical research* 34, 33–45.

Pritchard, D., (2010), “Knowledge and understanding”, in *The nature and value of knowledge: three investigations*, Pritchard, D., Millar, A. and Haddock, A., (eds.), 3–90, Oxford: Oxford University Press.

Psillos, S., (1999), *Scientific Realism: How Science Tracks Truth*, London-New York: Routledge.

Rasmussen, D. and Eliasmith, C., (2011), “A neural model of rule generalization in inductive reasoning”, *Topics in Cognitive Science*, 3, 140–53.

Räz, T., and Beisbart, C. (2022). “The Importance of Understanding Deep Learning”, *Erkenntnis*, 1–18.

Razzano, M., (2021), *Ascoltare il cosmo. Le frontiere dell’astrofisica dai neutrini alle onde gravitazionali*, Roma: Carocci.

Reardon, S., (2024), “Mini-colon and brain ‘organoids’ shed light on cancer and other diseases”, *Nature news* (24 April 2024).

Rescher, N., (1962), “The Stochastic Revolution and the Nature of Scientific Explanation”, *Synthese*, 14, 200–215.

Rice, C. C. (2016). “Factive scientific understanding without accurate representation”, *Biol. Philos*., 31, 81–102.

Riggs, W., (2008), “The Value Turn in Epistemology”, in *New Waves in Epistemology*, Pritchard, D. and Hendricks, V., (eds.), New York: Palgrave MacMillan, pp. 300–323

Rohwer, F., and Edwards, R. (2002). “The Phage Proteomic Tree: A Genome-Based Taxonomy for Phage”. *Journal of Bacteriology*, 184, 4529–35.

Rosenblueth, A., and Wiener, N., (1945), “The Role of Models in Science”, *Philosophy of Science*, 12, 4, 316–321.

Rovelli, C., (2019), *Il successo empirico della relatività generale e le sue implicazioni filosofiche per la comprensione della natura dello spazio e del tempo*, Roma: Bardiedizioni.

Rovelli, C., (2023), "The Relational Interpretation", in *The Oxford Handbook of the History of Quantum Interpretations*, Freire, O. Jr., Bacciagaluppi, G., Darrigol, O., Hartz, T., Joas, C., Kojevnikov, A. and Pessoa, O., Jr, (eds.), 1055–71, Oxford: Oxford University Press.

Rumelhart, D., E., McClelland, J., L., and the PDP Research Group, (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, Mass: The MIT Press.

Rudin, C., and Radin, J., (2009), "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition". *Harvard Data Science Review*, 1.2, 1–10.

Sainsbury, R. M., (1997), "Easy possibilities", *Philosophy and phenomenological research*, 57(4), 907–19.

Salmon, W., (1989), *Four Decades of Scientific Explanation*. Pittsburgh Press: University of Pittsburgh.

Salmon, W., (1992), *40 anni di spiegazione scientifica. Scienza e filosofia 1948-1987*. Padova: Muzzio.

Salmon, W., (1994), "Causality without Counterfactuals", *Philosophy of Science*, 61, 2, 297–312.

Scheffler, I., (1972), *The Anatomy of Inquiry*. Indianapolis: Hackett Publishing Company.

Scheffler, I., (2005), *Explanation and Understanding*. New York: Dover Publications.

Scheffler, I., (2015), "Understanding, Explaining, and Scientific Realism". *Philosophy of Science*, 82, 5, 778–89.

Schmueli, G., (2010), "To explain or to predict?", *Statistical Science*, 25(3), 289–310.

Scholz, R.W. and Tietje, O., (2002), "Embedded Case Study Methods: Integrating Quantitative and Qualitative Knowledge", *Thousand Oaks, CA: Sage Publications*.

Schrödinger, E., (1928), *Collected Papers on Wave Mechanics*, London: Blackie & Son.

Schrödinger, E., [1954] (1996), *Nature and the Greeks*, Cambridge: Cambridge University Press.

Sharma, A., and Kappeler, K., (2020), "Understanding Synthesis: Approaches in Molecular Nanotechnology". *NanoEthics*, 14, 2, 143–56.

Shrader-Frechette, K., (2014), *Method in Ecology: Strategies for Conservation*. Cambridge: Cambridge University Press.

Skelac, I,, and Jandrić, A., (2020), "Meaning as Use: From Wittgenstein to Google's Word2vec", in *Guide to Deep Learning Basics. Logical, Historical and Philosophical Perspective*, edited by Skansi, S., (ed.), 41–53, Berlin: Springer.

Smolensky, P., (1991), "Connectionism, constituency and the language of thought", in *Connectionism: Debates on psychological explanation* (Vol. 2), MacDonald, C., and MacDonald, G. (eds.), 164–198. Oxford: Blackwell.

Smolensky, P., (1995), "Reply: Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture", in *Connectionism: Debates on psychological explanation* (Vol. 2), MacDonald, Cynthia, and MacDonald, Graham, (eds.), 223–90. Oxford: Blackwell.

Sokol, J., (2024), "How The Ancient Art of Eclipse Prediction Became an Exact Science", *Quanta Magazine*, April 5, 2024, https://www.quantamagazine.org/how-the-ancient-art-of-eclipse-prediction-became-an-exact-science-20240405/ (last view 9/4/2024).

Sosa, E., (1999), "How to defeat opposition to Moore", *Noûs*, 33, 141–153.

Steel, D., (2010), *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.

Stern, D., G., (1991), "Models of memory", *Philosophical Psychology*, 4(2), 203–17.

Stern, D., G., (2007), "The uses of Wittgenstein's beetle: Philosophical Investigations §293 and its interpreters", in *Wittgenstein and his Interpreters: Essays in Memory of Gordon Baker*, Kahane, G., Kanterian, E. and Kuusela, O., (eds.), 248–68. Malden: Blackwell.

Stern, D., (2011), "Private Language", in *Oxford Handbook of Wittgenstein*, Kuusela, Os., and McGinn, M., Oxford: Oxford University Press.

Strevens, M., (2008), *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Strevens, M., (2013), "The Explanatory Role of Conserved Quantities". *Noûs*, 47, 2, 267-96.

Strueber, K., R., (2006), *Rediscovering empathy: agency, folk psychology, and the human sciences*, Cambridge, Mass.: MIT Press.

Strueber, K., R., (2012), „Understanding Versus Explanation? How to Think about the Distinction between the Human and the Natural Sciences", *Inquiry*, 55(1), 17–32.

Sullivan, E., (2022), "Understanding from Machine Learning Models", *Br. J. Philos. Sci.* 73(1), 109-133.

Sullivan, E., (2023), "How Values Shape the Machine Learning Opacity Problem", in. *Scientific Understanding and Representation: Modeling in the Physical Sciences*, Khalifa, K., Lawler, I., Shech, E. (eds.), Routledge, London.

Swoyer, C., (2008), "The Nature of Natural Laws". In Psillos, S., Curd, M., and Clarke, C. (eds.), *The Routledge Companion to Philosophy of Science*. New York: Routledge, 373-85.

Tamir, M. and Shech, E., (2023), "Understanding from Deep Learning Models in Context", in. *Scientific Understanding and Representation: Modeling in the Physical Sciences*, Khalifa, K., Lawler, I., Shech, E. (eds.), Routledge, London.

Tietje, O. (1998). "Explanation and Understanding Revisited". *Journal for General Philosophy of Science*, 29, 1, 129–53.

Tourlet, S., Radjasandirane, R., Diharce, J. and de Brevern, A., G., (2023), "AlphaFold2 Update and Perspectives", *BioMedInformatics*, 2023, 3(2), 378–90.

Turina, P., Fariselli, P. and Capriotti, E., (2023), "K-Pro: Kinetics Data on Proteins and Mutants", *Journal of Molecular Biology*, 435, 168245–51.

van Fraassen, B.C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

van Fraassen, B.C. (1989). *Laws and Symmetry*. Oxford: Oxford University Press.

van Fraassen, B.C. (2008). *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

van Fraassen, B.C. (2010). *The Empirical Stance*. New Haven: Yale University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., (2017), "Attention is all you need", *31st Conference on Neural Information Processing Systems* (NIPS 2017), Long Beach, CA, USA, 1–15.

Vellejos-Baccelliere, G. and Vecchi, D., (2023), "Searching for Protein Folding Mechanisms: On the Insoluble Contrast Between Thermodynamic and Kinetic Explanatory Approaches", in *New Mechanism. Explanation, Emergence and Reduction*, Cordovil, J., L., Santos, G. and Vecchi, D., (eds.), Netherlands: Springer, 109–37.

Von Wright, G.H., (1971). *Explanation and Understanding*. London: Routledge.

Wang, Alex *et al*. 2019. 'SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems', in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, edited by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox. E., and Garnett, R., 1–15.

Watts, D. J., Strogatz, S. H., (1998), "Collective dynamics of 'small-world' networks. Nature 393, 440–44.

Weisberg, M., (2007), "Three Kinds of Idealization". *Journal of Philosophy*, 104, 12, 639–59.

Weisberg, M., (2013), *Simulation and Similarity: Using Models to Understand the World*, Oxford: Oxford University Press.

Wiener, N. (1948), *Cybernetics, or control and communication in the animal and the machine*, Cambridge (Mass): The MIT Press.

Wilkenfeld, D. A., Plunkett, D. and Lombrozo, T., (2016), "Depth and deference: when and why we attribute understanding." *Philosophical studies* 173(2), 373–93.

Williamson, T., (2000), *Knowledge and its limits*, Oxford: Oxford University Press.

Wittgenstein, L., [1921] (1961), *Tractatus logico-philosophicus*, edited by Pears, David, F., and Brian, F., McGuinness. New York: Humanities Press.

Wittgenstein, L., (2009), *Philosophical investigations*, 4th edition, Hacker, P., M. S., and Schulte, J., (eds.), Oxford: Wiley-Blackwell.

Wittgenstein, L., [1914-1916] (1969), *Notebooks 1914-1916*, von Wright, Georg, H., and Anscombe, G., E. M., Anscombe, New York: Harper Torchbooks.

Wittgenstein, L., [1933–5] 1958, *The Blue and Brown books*. New York: Harper and Row.

Wittgenstein, L., (1978), *Philosophical grammar*dited by Rhees, R., Kenny, ., Berkeley: University of California Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Wray, K.B. (2002). *Transitive Reduction: A Theory of the Essence of Mathematical Ontology*. Oxford: Oxford University Press.

Wray, K.B. (2006). "Models and Explanation". *Synthese*, 152, 359–73.

Zacharias, N. (2018). "Cellular Responses to DNA Damage: One Signal, Multiple Pathways". *Trends in Cell Biology*, 28, 6, 426–38.

Zagzebski, L. T., (1996), *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*, New York: Cambridge University Press.

Zagzebski, L. T., (2001), "Recovering Understanding", in Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue, Matthias Steup (ed.), New York: Oxford University Press, pp. 235–252.