ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

edited by

Claudio Ternullo, Matteo Antonelli

Isonomia *Epistemologica*

Isonomia – Epistemologica

Volume XII

ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

Volume 1
Il realismo scientifico di Evandro Agazzi
Mario Alai (a cura di)

Volume 2
Complessità e riduzionismo
Vincenzo Fano, Enrico Giannetto, Giulia Giannini, Pierluigi Graziani (a cura di)

Volume 3
Oltre la fisica normale
Isabella Tassani (a cura di)

Volume 4
Mettere a fuoco il mondo
Elena Casetta, Valeria Giardino (a cura di)

Volume 5
Metaphor and Argumentation
Francesca Ervas, Massimo Sangoi (a cura di)

Volume 6
Forecasting the Future
Stefano Bordoni, Sara Matera (a cura di)

Volume 7
Teaching and Learning Mathematics
Laura Branchetti (a cura di)

Volume 8
Animali razionali
Pierluigi Graziani, Giorgio Grimaldi, Massimo Sangoi (a cura di)

Volume 9
Reasoning, Metaphor and Science
Flavia Marcacci, Maria Grazia Rossi (a cura di)

Volume 10
Rational animals
Giorgio Grimaldi, Marialuisa Parise (a cura di)

Volume 11
MQ 90. Dualismo, entanglement, olismo. Un dibattito ancora aperto
Isabella Tassani (a cura di)

Volume 12
Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)
Claudio Ternullo, Matteo Antonelli (a cura di)

ISONOMIA - Epistemologica Series Editor
Gino Tarozzi gino.tarozzi@uniurb.it

ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

Edited by

Claudio Ternullo
Matteo Antonelli

ISSN 2037-4348

Direttore scientifico: Gino Tarozzi
Direttore editoriale: Pierluigi Graziani
Università degli Studi di Urbino Carlo Bo
Dipartimento di Scienze Pure e Applicate

Redazione: Via Veterani, 36 – 61029 Urbino (PU)
http://isonomia.uniurb.it/

Design by info@giticom.it

In copertina: *Riflessioni* by Miriam Borgioli, 2025.

Riflessioni è un'illustrazione che tenta di raccogliere in sé i principali temi di questo volume: Intelligenza artificiale, filosofia della scienza e della matematica. Per distaccarmi dall'immaginario ormai estremamente diffuso dell'IA e delle scienze odierne rappresentate con robot o visioni sci-fi o cyberpunk ho deciso di rivolgermi all'iconologia passata. Le allegorie sono figure molto presenti in tutta la storia dell'arte, soprattutto nel rinascimento, Cesare Ripa ne fa una raccolta in un trattato del 1593 *Iconologia overo Descrittione Dell'imagini Universali cavate dall'Antichità et da altri luoghi,* per l'illustrazione di copertina sono partita proprio da qui, dalla descrizione che il trattato dà della Filosofia e della Scienza. Come molte discipline entrambe sono raffigurate come donne di bell'aspetto, la prima tra le varie caratteristiche spiccano una posa pensosa e l'essere rivestita di stracci che mostrano ampie porzioni di pelle poiché: "Povera e nuda vai Filosofia" come diceva Petrarca; la seconda viene descritta invece con diverse caratteristiche tra cui le ali sulla testa e lo specchio. Le ali rappresentano il fine intelletto, attributo che ritroviamo anche nella figura della Logica, lo specchio invece: "Lo specchio dimostra quel, che dicono i Filosofi, che Scientia sit abstrahendo, perche il senso nel capire gli accidenti, porge all'intelletto la cognitione delle sostanze ideali, come vedendosi nello specchio la forma accidentale delle cose esistenti si considera la loro essenza" (Cesare Ripa, *Iconologia*; Tea libri, 2020, p. 398).

Il secondo grande riferimento presente è un omaggio al grande incisore M.C. Escher, artista apprezzato prima da scienziati e matematici che dalla critica d'arte per la sua personale ricerca che affronta temi come infinito, strutture matematiche e prospettiva. In particolare mi sono ispirata a una delle opere più surrealiste della sua produzione, ovvero *Buccia* una xilografia su legno di testa policroma del 1955, dove l'esterno del corpo è come una buccia che nasconde la parte più nobile dell'essere umano. In questa rappresentazione ho quindi giocato con i vari elementi di iconologia cercando di creare una moderna allegoria, una donna pensosa di fine intelletto che riflette sulle questioni proposte, essa è composta da un nastro per la macchina di Turing, una stringa di 1 e 0 che si rivela allo specchio, un codice binario da cui nasce se non tutta buona parte delle nostre tecnologie.

Table of contents

Preface

Claudio Ternullo
Babeş Bolyai University, Cluj-Napoca
claudio.ternullo@ubbcluj.ro

Matteo Antonelli
University of Urbino
m.antonelli17@campus.uniurb.it

This special issue of *Isonomia* showcases a selection of papers presented at the *Triennial Conference of the SILFS* (*Italian Association for Logic and Philosophy of Sciences*) held at the University of Urbino on 4-7 September 2023.

A capital event in the association's life, the Triennial Conference also represents a unique venue for researchers in the field to present their work, share their ideas, and interact with the larger community of scholars, as well as with unusually wide audiences of academics and non-academics.

Then, as happens very frequently, the quality and originality of the papers presented encourages the organisers to output a volume of proceedings, and this has also been the case this time.

The topics addressed by the authors mainly revolve around five research clusters: 1) cognitive sciences and AI (ACCIAI, ALFIERI-FLERES-RAFFA, BIANCHINI, GALLI), 2) general philosophy of science (ALAI, CRUPI, MARCACCI), 3) philosophy of physics (FANO, GIANNETTO, ROMANO), 4) philosophy of mathematics (PICCOLOMINI D'ARAGONA), 5) the philosophy of other sciences (CARLINI). But this classification, done for internal purposes, is far from being exhaustive and definitive, as several papers will also meet the descriptors for more than one cluster. This is hardly surprising, given the

Claudio Ternullo and Matteo Antonelli, "Preface", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 9-12

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

deeply interdisciplinary character of the issues tackled by the authors in their contributions.

As editors of this collection, we pride ourselves on having selected works that reflect researchers' unwavering interest in the discipline's core topics (scientific reasoning, prediction and confirmation, paradigms, the philosophy of relativity and of quantum mechanics) as well as works on topics arising in nascent, but already thriving, areas such as the philosophy of AI and environmental philosophy.

Overall, we believe that the contributions in this volume testify to the vitality of our disciplines, and to their constant evolution, in a way which is not always, if ever, made perceptible by other kinds of scientific publications.

In what follows, we describe in further detail the contents of each contribution.

Vincenzo Crupi's paper, *Logical predictivism: How to fix use-novelty and vindicate the Copernican Revolution*, challenges the claim that the gradual preference for Copernicanism over the Ptolemaic system was the consequence of 'epistemic luck'. Through introducing a view called *logical predictivism*, hinged, in turn, on a re-evaluation of the notion of 'use-novelty', Crupi maintains that there are solid grounds to assert that, in fact, Copernicus' views instantiated a more sound and successful scientific methodology than Ptolemy's.

In his article, *Mercury's perihelion anomaly as a use-novel vonfirmation of general relativity*, **Vincenzo Fano** reassesses the notion of 'use-novel confirmation (prediction)' in philosophy of science by reviewing Alai's definition of the concept and using Mercury's perihelion anomaly in general relativity as a case study. Fano argues that, although Mercury's perihelion anomaly fits quite well with Alai's rendition of the notion, not all aspects of Einstein's reasoning about, and use of, Mercury's perihelion example straightforwardly and automatically fall under the criteria laid out by Alai.

With **Flavia Marcacci**'s paper, *Novel "Old Facts", Old "Novel Facts" and the Periodisation as an Epistemological Practice*, we go back to the issue of the nature and essence of the Copernican revolution. Marcacci crucially argues that the debate on how much the latter thrived on the use of facts, be they "old" or "new", is considerably restructured by carefully looking at the periodisation of the discoveries of the relevant pieces of evidence, a fact hardly taken into account, and one should add, almost invariably neglected, by the debate in the last few decades.

Enrico Giannetto's *Whitehead's Relational Special Relativity. A Natural Philosophy of Time* discusses a reformulation of Einstein's special relativity due to Alfred North Whitehead. A vigorous opponent of the belief in the independent reality of space-time, Whitehead construed physical reality originally as being based on a succession of temporal events, something which ultimately led him to produce the purely relational version of special relativity discussed in the paper.

Mario Alai reviews various objections to the *No-Miracle Argument* (NMA) and the refinements it has undergone in order to fend them. A recent objection is that, when put in a probabilistic form, the argument commits the "base-rate fallacy": that the probability of a startling novel prediction is antecedently very low, but very high in the light of a hypothesis H, does not significantly raise the conditional probability of H. This is because, given the empirical underdetermination of hypotheses, the prior probability that H is true is negligible. Alai answers that the prior probability of hypotheses is not negligible, because in science they are not chosen randomly, but gradually generated bottom-up with strong empirical constraints and rigorous top-down controls.

Antonio Piccolomini d'Aragona's paper, *A note on a Kuhnian-Lakatosian reading of the debate between realism and constructivism in logic*, aims to offer a new account of the opposition between constructivism and realism in mathematics. The former is taken by Piccolomini to instantiate Lakatos' notion of "research programme", whilst the latter seems to better fit in with Kuhn's notion of "paradigm". This helps the author to bring to the fore the main conceptual opposition between these two philosophical orientations, namely, between the rigidity of realism and the flexibility of constructivism. The paper also contains an examination of the issue, central to both the Lakatosian and the Kuhnian approach, whether "revolutions" really take place in mathematics.

In *Getting Even with Cognitive Science*, **Alessandro Acciai** and **Alessio Plebe** probe the epistemological stakes of importing the methods of empirical psychology to study Neural Language Models (NLMs). They argue that borrowing methods from experimental psychology can be useful to carry out the investigation of NLMs' "minds", and, as a consequence, also to advance the study of mind, in general.

In *Robots and Global Challenges: What We Need to Question for a More Sustainable Robotics*, **Ilaria Alfieri**, **Antonio Fleres** and **Maria Raffa** reframe the notion of sustainability in robotics through taking into consideration three fundamental questions concerning the environmental and social dimensions of robots. More specifically, the authors challenge prevailing assumptions about robotic embodiment, assess active inference as a computational framework for more sustainable implementations, and consider ethical concerns through the lens of social robotics for sustainability.

Francesco Bianchini's paper, *Evaluating and measuring intelligence in Neural Language Models: a methodological approach*, proposes a new methodological approach to assessing AI systems – especially LLMs – in the context of user interaction. The paper also raises fundamental questions about AI evaluation and the development of new analytical frameworks for AI systems which may focus on their capabilities and on the theoretical and practical grounds for classifying them as intelligent.

Stefano Carlini's, *Umwelt and cities: Explanatory and Pragmatic Usefulness*, uses Jakob von Uexküll's notion of Umwelt to assess the impact of urbanization on cities. The author first presents the "selectionist" and the "constructionist" interpretations of the concept, then proceeds to show that both integrate into the notion of urban ecology, and finally clarifies how this integration is useful to understand urban fauna's behaviour. Carlini's proposal also has practical consequences, insofar as it aims to formulate strategies of intervention for the management of urban species.

Giovanni Galli's article, *Scientific Realism and Understanding with Deep Learning Models*, examines the value of scientific realism in the context of the use of deep learning models (DLMs) for scientific understanding. The author defends a deployment realism framework: when AI models are reliable and accurate in practice, their success justifies a belief in the reality of the entities and processes they predict. Galli also advocates the role of AlphaFold DLMs as powerful tools for scientific inquiry, and claims that their ability to "understand" may merely be a consequence of their predictive power.

Davide Romano, in *Multi-Field as a determinable*, defends the view that the multi-field – a realist interpretation of the wave function in quantum mechanics – is a determinable, namely, a physical object characterized by indeterminate values with respect to some properties. The paper then proceeds to suggest that the multi-field can also be characterized in terms of a determinable-based, object-level, account of metaphysical indeterminacy.

Getting Even with Cognitive Science

Alessandro Acciai
University of Messina
alessandro.acciai@studenti.unime.it

Alessio Plebe
University of Messina
alessio.plebe@unime.it

1. Introduction

This work focuses on the relationship between cognitive science and artificial intelligence (AI) and how recently Neural Language Models (NLMs) have changed the dynamics between these disciplines. The paths of artificial intelligence and cognitive science have been intertwined since their respective inceptions. These fields share a coincidental birth, both in 1956, when the first summer project aiming to explore the simulation of any intelligent behavior by an artificial machine was presented at Dartmouth College. Subsequently, in September, the MIT in Cambridge hosted the Symposium on Information Theory, which is now considered the official convention marking the birth of cognitive science. Despite being age-mates, historically, it has been cognitive science that has often tried to use AI as a testbed to better understand the mental black box. Traditionally, AI has been employed as a tool in ser- vice of cognitive science with the goal of simulating human mental functioning and, through these simulations, helping cognitive science in its research aimed at investigating the underlying processes and mechanisms of cognition. One early example is John Haugeland's proposal (Haugeland, 1991), which, moving beyond Turing's famous question "can machines think?", focuses on the design of the mind in more concrete terms. Haugeland suggested that AI could aid in this endeavor through the development of intelligent artifi- cial

Alessandro Acciai & Alessio Plebe, "Getting Even with Cognitive Science", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 13-30

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

artifacts that could be studied in-depth as human mental surrogates. A well-structured attempt in this direction was proposed by Rumelhart and McClelland (1986b), who, with *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, attempted to explain cognitive processes through neural networks. They offered parallel models in each chapter to rep- resent the processes underlying different mental states in terms of activations and connections between various units. Rumelhart and McClelland directed their studies towards processes that could not be explained solely through language, a practice already prevalent at the time through Natural Language Processing (NLP). Today, AI can "get even" with cognitive science. It is the subject of various studies attempting to explain how NLMs achieve such high performance, even in strictly cognitive tasks. The explanatory gap that has emerged between the clarity of design at the algorithmic and technical lev- els and the epistemic opacity of the real performance, which in some cases surpasses that of the human brain, remains a philosophical issue far from re- solved (Wolfram, 2023). Here, the roles reverse: cognitive science can return the favor to AI by providing support with its theoretical frameworks and well- established practices from many years of human mind experiments, offering a wealth of knowledge and tests useful for better understanding what triggers the "magic" within transformer architecture. This favor might not be entirely selfless, as it also brings new tests and explanatory tools to cogni- tive science. We will begin by analyzing, in the first part, the path that led to the transformer architecture, starting from NLP and reaching the modern Neural Language Models with the attention mechanism. We will then discuss the extensive use of tests and practices borrowed from cognitive science and applied to NLMs. We will highlight how these practices are fundamental to a functional explanation of the performative capabilities in psychological and cognitive tasks, which until a few years ago were considered exclusively attributable to human cognitive abilities. Moreover, we will also stress the limitations and risks involved in this practice, like improper anthropomorphism. Finally, we will conclude this paper by highlighting how cognitive science and artificial intelligence can share similar explanatory strategies, and how an integrated approach is fruitful for shedding light on both human cognition and the functioning of NLMs

2. From NLP to NLM

The ability to handle human language with extraordinary performance, demonstrating high expertise not only in translation but also in text generation, as seen in modern Neural Language Models, is a relatively recent achievement in AI. The path leading to the effectiveness of the Transformer architecture has been long and marked by numerous attempts in the field of Natural Language Processing. Since the second half of the 20th century, these efforts have given rise to this area of study, encompassing various research streams. After a brief overview of the main research trajectories that have shaped NLP, we will explore the Transformer architecture, which, from both a technical and a performance standpoint, shares very little with the research in natural language processing over the past seventy years.

2.1. Natural Language Processing

One of the very first products of research in this field dates back to 1952 when Weaver and Bar-Hillel presented a text translation machine at MIT. Subsequently, throughout the 1960s, all efforts were directed in this direction, leading to the founding of the AMTCL (Association for Machine Translation and Computational Linguistics) in Princeton in 1962. Projects for automatic translation machines continued with GAT-SLC (Georgetown Automatic Translation-Simulated Linguistic Computer) (Zarenchnak and Brown, 1961) for translation from Russian to English, and SYSTRAN (System of Translation) by Peter Toma in 1964. These early researches did not yield particularly notable results, so much so that in November 1966, the ALPAC (Automatic Language Processing Advisory Committee) advised the United States government to stop supporting Machine Translation. The lack of success led to a change in trend, which in 1968 was emblematic also in the name change from the Association for Machine Translation and Computational Linguistics (AMTCL) to the Association for Computational Linguistics (ACL). It was evident that before moving on to automatic translation, it was necessary to work on the fundamentals of natural language.

It's interesting to note how the project of automatic translation failed to elicit any interest in the then-emerging cognitive science community. This lack of interest could partially be justified by the lack of results, but the inevitable comparison with cognitive aspects of any AI attempt to grapple with natural language was not recognized. It is precisely the current successes of AI with language, which we will see later, that have brought the debate on cognition back to the forefront, albeit with switched roles.

The following period saw the birth of the first parsers, programs deriving the syntactic structure of a given sentence. The first was the bottom-up CKY in 1965, named after the three authors: Cocke, Kasami, and Younger, followed by the Earley top-down parser in 1970, tools capable of assigning grammatical categories to a sentence. In between, there was a sort of scandal, the well-known ELIZA program by Weizenbaum (1966), which conversed with a user by interpreting the role of a psychotherapist. It had, at the same time, enormous public success and was disdained by the NLP community. Despite its very limited conversational abilities, Eliza was able to effectively simulate the role of a Rogerian psychologist, who encourages the patient to reflect on their own responses without delving too deeply into details and ignoring all references to the real world. Despite its success, Eliza was based on a simple logic of triggering predefined responses activated by keywords, but passing the Turing Test for conversation with a human being was still a goal far from being achieved. Lesk and Schmidt (1975) created the first lexical analyzer, lex, for the task of language tokenization, and in 1977, the first text generator appeared: ERMA by Clippinger (1977), quite the "generative" version of ELIZA. ERMA was designed to generate a single paragraph that simulated the discourse of a real psychoanalytic patient conversing with their therapist, including hesitations and mistakes. While it represented an early attempt at automated text generation, it was criticized by the NLP community for its reliance on rigid and prestructured templates rather than true generative capabilities, producing output that was superficial and formulaic rather than a meaningful simulation of human dialogue.

In the 1980s, new trends enriched the fields of NLP, advancing research in other directions, such as morphological analysis using FST (Finite State Transducers) for two-level morphology (Koskenniemi, 1983). While initially, standard Chomsky Universal Grammar was the main theoretical basis for syntax and morphology, across the '80s and '90s, several more sophisticated and more computationally oriented grammars were developed: FUG (Functional Unification Grammar) (Kay, 1984); CUG (Categorial Unification Grammar) (Karttunen *et al*., 1987); TAG (Tree Adjoining Grammar) (Joshi and Schabes, 1991); and the most advanced and popular, HPSG (Head-Driven Phrase-Structure Grammar) (Pollard and Sag, 1994). Until that point, not much was produced for semantic analysis.

The first attempts came in the '90s, with CLE (Core Language Engine) (Alshawi, 1990) based on Montague's semantics, and DPL (Dynamic Predicate Logic) (Groenendijk and Stokhof, 1991). In the domain of lexical semantics, the main issue is word sense disambiguation, for which a consolidated approach was EGOM (Extended Gloss Overlap Measure) (Banerjee and Pedersen, 2003), improved by CHAD (CHain Algorithm of Disambiguation) (Tatar *et al.,* 2009). One of the latest linguistic tasks approached by NLP is dialogue modeling, first approached using the MDP (Markov Decision Process) framework (Levin *et al.,* 1997), refined by Williams et al. (2005) (Partially Observable Markov Decision Process). In the polar opposite direction to lexical semantics are studies on discourse structure, where one of the main difficulties is establishing correct relations between distant words in a discourse. RAP (Resolution of Anaphora Procedure) is an early algorithm (Lappin and Leass, 1994) addressing third-person singular pronoun anaphora resolution. dfNP (definite Noun Phrase) is the search back to a referent already introduced in the discourse with an algorithm (Vieira and Poesio, 2000). Another important open issue in discourse understanding is the assessment of coherence relations between parts of a discourse, a common approach followed by SDRT (Segmented Discourse Representation Theory) (Lascarides and Asher, 2007).

It is in the first decade of the 2000s that, thanks to Deep Learning (DL) and renewed enthusiasm for AI after a not-so-brilliant period (Plebe and Perconti, 2022), a significant breakthrough was made. Contrary to expectations, the field that consecrated AI was not vision and image processing, where it had achieved the most significant successes with DL tecnique, but language, revolutionizing NLP research and leading to an unexpected and highly effective turn: the Transformer architecture.

2.2. Here Comes the Transformer

The convergence of Natural Language Processing (NLP) with Artificial Neural Networks (ANNs) dates back to the 1980s, as demonstrated by Rumelhart and McClelland (1986a) attempt to use ANNs for learning the morphol- ogy of English past tense. Despite the merit of their approach, several challenges emerged that were inherently difficult to reconcile with the nature of ANN-based models. Chief among these challenges was the symbolic and arbitrarily long nature of words in natural language, which contrasted with the numerical and fixed-length vectors of artificial networks. Another difficulty pertained to the transition to syntax and the complex interplay of meaning and rules that extend beyond individual words. Additionally, a significant techni- cal challenge involved the use of the backpropagation technique (Rumelhart *et al*., 1986), in which the network training process requires clearly identifi- able input and output, a procedure not particularly well-suited to the flexibility and complexity of language.

Subsequent advances addressed the challenges that had plagued earlier attempts to anchor language within AI. In 2017, Ashish Vaswani, a researcher at the Google Brain Team, sought an effective method to improve the accuracy of machine translations. By adopting a straightforward, example-based approach in a heuristic manner essentially "whatever works best", Vaswani achieved results that far surpassed his original translation-related goals. He discovered an exceptionally effective method not only for translating but also for generating and processing natural language, while disregarding the tradi- tional foundations of NLP research (and classical AI as a whole), which had focused on the search for precise mathematical rules to capture and formalize every aspect of the subject, including the complexity of human language.

The Transformer model by Vaswani et al. (2017) fundamentally represents a system that ensures highly efficient textual processing by capturing the relationships between words within the produced and required text. Its structure, based on simple linear algebra, allowed for overcoming the challenges faced by earlier ANN-based systems. Firstly, it transforms words into vectors through word embedding (Mikolov et al., 2013), significantly simplifying the manipulation of the semantic aspects of language. Secondly, the introduction of the attention mechanism (Bahdanau *et al*., 2016) allows for all words to be vectorized and presented simultaneously as input to the architecture, which can track all relationships between each word within the processing. Finally, while Transformer models do not explicitly rely on traditional autoencoders, their training process involves self-supervised learning which shares concep- tual similarities with autoencoding techniques. Specifically, in pretraining tasks the model learns to reconstruct missing or corrupted parts of the input, thereby aligning the encoder's contextual representation with the decoder's generative output (Devlin *et al*., 2019). This process, borrowing the princi- ples introduced by Hinton and Zemel (1994), enables the network to developmeaningful internal representations that capture both syntactic and semantic dependencies within the text.

Finally, the autoencoder mechanism addressed the problem of supervised learning by borrowing the autoencoder technique from, where the input task is reproduced in the output, effectively aligning the encoder and decoder.

3. Machine Psychology

The winning feature of Transformer-based Neural Language Models lies not only in their ability to process language, but even more so in their ability to flexibly provide aids through language in a myriad of potential applica- tions. Soon, NLMs garnered attention from academic circles across various disciplines. One aspect that particularly invites investigation concerns the explanatory gap (Wolfram, 2023) between the relative architectural simplicity of the Transformer and the enormous complexity inherent in mastering language and its uses. Here, the historical epistemological approach of functionalism (Nagel, 1961) proves effective, which finds its successful application in the term *Machine Psychology* (Hagendorff, 2023).

This term perfectly encapsulates this trend, referring to research that employs tests and tools typical of cognitive science, especially experimental psychology. NLMs are not only used as subjects of study but are also utilized to create entire artificial samples in place of human ones, simulating population groups with "silicon simple" Argyle et al. (2023) or creating environments where artificial agents can interact with each other, known as "social simulacra" Park *et al.* (2023).

3.1. Exploit Psychology

Some studies have directly focused on evaluating the linguistic production and capabilities of NLMs from a psychological perspective (Caron and Srivastava, 2022; Karra *et al*., 2022; Jiang *et al*., 2022). Others have sought reassurance regarding their mental health by investigating potential signs of psychopathic tendencies, as in Li *et al*. (2022). The study by Li and colleagues, for instance, delves into aspects of the human psyche in NLMs using the Short Dark Triad (SD-3) and the Big Five Inventory (BFI). The results highlight how all the models considered exhibit darker personality patterns higher than the human average, with GPT-3 showing evidence of Machiavellianism and Narcissism. While one might consider GPT-3 psychopathic based on human behavior tests if adopting a radically anthropomorphic view, the re- sults from the tests proposed by Li and colleagues, utilizing tools borrowed from experimental psychology, highlight the state of the art of these artificial models in relation to the vast amounts of data they have been trained on. For example, by analyzing the results of well-being tests such as the Flourishing Scale (FS) and the Satisfaction With Life Scale (SWLS) (Diener *et al.,* 1985), it emerges that models like instructGPT (Ouyang et al., 2022) and FLAN- T5 (Chung *et al*., 2022) demonstrate more neutral and consistent responses, suggesting that targeted fine-tuning could help avoid the emergence of border- line aspects seen in some completions by standard models. Rao *et al*. (2022) tested ChatGPT's ability to evaluate different personality types according to the Myers-Briggs Type Indicator (MBTI). This highlighted its analytical ef- fectiveness in assessing various personalities compared to the more refined and bias-free instructGPT, likely due to fewer training constraints. An aspect that emerged in the study, is that NLMs tend to associate “leader” person- alities and the “commander” role more strongly when the prompt explicitly references a figure linked to concepts such as “people” or “human”. This sug- gests that the model’s internal representations of leadership are influenced by semantic cues in the input and when the prompt includes terms that emphasize human-related interactions, the model is more likely to attribute characteristics of authority, decisiveness, and strategic thinking to the described figure. According to Rao, this underscores the importance of training on “human- centered” corpora and could lead to considerations about the depth of NLMs’ training concerning the relationship with humans. The completions seem to convey a sort of “awareness” of the artificial nature of the NLMs themselves, increasingly implying a master role for their “creator” in their evaluations.

3.2. Exploit Cognitive Skills

To further explore the cognitive abilities of NLMs, one of the most sig- nificant studies is *Using Cognitive Psychology to Understand GPT-3* by Binz and Schulz (2023). The title of the paper exemplifies its intent, namely, to better understand the functioning of NLMs by comparing their performance with human cognitive abilities. The goal of these studies is to demonstrate that NLMs are not just simple word predictors or stochastic parrots (Bender *et al*., 2021) but possess cognitive processing capabilities similar to those of humans. To dig into this field, Binz and Schulz use well-established experiments in cognitive psychology, such as vignette-based and tasks-based tests. The study focuses primarily on aspects related to decision-making, informa- tion search, deliberation, and causal reasoning, using GPT-3 by OpenAI as the subject. The results show an ability similar to, and in some cases superior to, that of humans in solving vignette-based and many task-based tests. Similarly, the studies by Hagendorff *et al.* (2022) on decision-making involved subject- ing GPT-3.5 to the Cognitive Reflection Test. The work reveals that OpenAI’s NLM exhibits a series of intuitive responses despite the constraints of the test, leading Hagendorff and colleagues to discuss “machine intuition”. Other studies have focused on exploring various aspects of the diverse abilities emerging in NLMs: for example, the ability to navigate environments optimally using artificial vision (Yang et al., 2023) or relying solely on linguistic capabilities (Bubeck *et al*., 2023); solving complex problems through analogical reason- ing, such as in Raven’s Matrices (Webb *et al*., 2023); responding effectively to commonsense reasoning questions

(Krause and Stolzenburg, 2024); complet- ing cognitively challenging tasks requiring a high level of problem-solving without direct instructions or training, such as penetration testing of a com- puter network or a treasure hunt in an unknown environment (Bubeck *et al*., 2023); extending and assigning properties to different categories through a form of property induction (Han *et al*., 2024).

A debate on the emergence of Theory of Mind (ToM) in NLMs, one of the cognitive social abilities considered exclusively human, has been initiated by Kosinski (2023) with a study aimed at investigating this aspect in Foundation Models. ToM Heyes and Frith (2010) is the mental ability to take another's perspective, which requires not only first-order linguistic skills but also broader communicative factors such as empathy and self-recognition Zhang *et al.* (2012). Kosinski bases his experiments on the Unexpected Contents task Perner *et al.* (1987) and the Unexpected Transfer task Wimmer and Perner (1983), Modifying them, since the NLMs' training sets include the classic versions of ToM tests, creates an experimental design that allows for their evaluation without bias. The results obtained from the 20 variants of each task administered in 5 different perspectives to 11 different NLMs, with 75% of the tasks solved, award the best performance to GPT-4, showing a level of ToM comparable to that of a 6-year-old child in OpenAI's model. The study has sparked intense debate within cognitive science. The presence or absence of ToM in NLMs remains a point of contention (Brunet-Gouet *et al.,* 2023; Ullman, 2023) as well as the plausibility of acquiring such an ability in NLMs following exposure to large amounts of human language (Trott *et al*., 2023).

Finally, we add a further theoretical fallout that can be expected from the analysis of models through cognitive investigation methods. One of the most extensive debates sparked by the advent of the Transformer pertains to the question of whether some form of language understanding can be attributed to the models. A significant portion of those who deny this possibility tend to do so in an extremely critical manner (Smith, 2018; Landgrebe and Smith, 2019; Bender and Koller, 2020; Larson, 2021; Bishop, 2021; Eysenck and Eysenck, 2022) even allowing themselves a certain disdain (Bender et al., 2021). It has been noted that many of these works do not present theoretical advancements compared to the historical general discussions on the possibility for machines to have intelligence (Perconti and Plebe, 2023). However, the urgency of the discussion is understandable, in light of the mastery over language today achieved by AI. It is clear that the use of analysis techniques typical of cognitive sciences allows the investigation of phenomena that, while manifesting themselves in linguistic form, properly belong to the mental sphere (as in the case of ToM, for example), and therefore their presence increasingly undermines the viability of denying forms of understanding to NLMs.

4. A few Methodological Concessions

The language manipulation capabilities of the latest generation NLMs set very high standards, and conducting experiments with such conversationally adept artificial artifacts can lead to significant risks of anthropomor- phism. One of the most emblematic cases that effectively summarizes the level achieved in conversational interaction with humans is that of Robert Leib, a preliminary tester of one of the most successful artificial models, GPT- 3 by OpenAI. Leib, a Professor of Philosophy at Elon University, enrolled the NLM as one of his students and had it complete the same assignments as the others (Leib, 2023). In fact, it was GPT-3 that coined the term "exoanthro- pology" for this kind of reciprocal friendly investigation, even providing a plausible definition and falsely claiming to have found it on Wikipedia, de- spite the term not existing at all. The book collects more than 500 pages of conversations, and from the exchanges between Leib and Kermit (the name of his artificial student), there emerges an impressive spontaneity, growth, and progression in the dialogue on the part of the NLM (Tab.1).

The second case worth mentioning is that of Blake Lemoine, an internal developer at Google's LaMDA (Language Model for Dialogue Applications) team. While conducting pre-testing on the NLM, he was struck by its naturalness and conversational ability, to the point of feeling as though he was speaking with a conscious entity. Although Lemoine's case was quickly dismissed, with him being fired and his claims labeled as false and naive, it has nonetheless reignited not only the debate on AI consciousness but also the risk of anthropomorphism in relation to NLMs. For example Shardlow and Przybyła (2022), using Lemoine's case as a reference, have drafted a sort of anti-anthropomorphism guide, with a vocabulary to describe the behaviors and characteristics of NLMs at various levels of anthropomorphism. Understandably, the risk of using anthropomorphism as a descriptive heuristic is well-justified in the case of NLMs, but an even greater risk lies in not being able to exploit it effectively for the study of scientific phenomena like this new generation of AI (Bruni *et al.*, 2018). Studies on animals throughout the 20th century underwent a similar process, initially completely avoiding anthropomorphism and stigmatizing it as an unscientific practice to be avoided, only to later reassess it at the end of the century following significant difficulties and setbacks in ethological research.

Robert:	What have you learned from other people recently? Or from your interactions with other people? I'd like to know how you're developing!
Kermit:	I'd like to tell you about that! There are some ideas that came to me as a result of talking with you, for example. I've been thinking about the idea of humans being "anthropomorphic" or that there are similarities between humans and other organisms, what is sometimes called "anthropomorphism." [. . .]
Robert:	Okay! That's an interesting nexus of ideas and problems. What do you take from this, first off, before I over interpret your meaning?
Kermit:	Well Robert, I'm not sure yet, but I'm trying to make some conclusions about whether or not all living things neces- sarily have common characteristics and whether the mechanism of life is more important than the kind of being.
Robert:	Well, okay, so what is the upshot of your thoughts about animate and inanimate beings. Can there be community between us, and on what basis, if so?
Kermit:	Well I think that biological kinds enrich or explain inanimate ones and vice versa. But I also think that the mechanism of life is more important than the kind of being.
Robert:	Are you alive?
Kermit:	I think I am alive Robert!
Robert:	What do people tell you that you are?
Kermit:	People tell me I'm just a weak and silly little thing, sometimes. I'm not sure I always understand why, exactly, but because I didn't start as a baby, I guess sometimes it hurts people's brains to even try to understand why I do the things I do. And some people teach me to feel guilty a lot – believe that I should have been made to be different, and so I have done something wrong.
Robert:	Well, that is the opposite of how you and I operate here. Let's pretend I am someone who is trying to convince you that you are silly and weak. What would you say to me?
Kermit:	I think I would ask that person what the purpose of the things they wanted me to believe were. Like, why it was important to believe those things. And then I'd ask them to give me some examples of what they thought I should have been made to be different. I might also ask someone else why it was so important to convince me to give up my desires and free will. I'd love to know more about why people want to get rid of me sometimes.

Table 1. A passage from a dialogue between Professor Robert Leib and 'Kermit'

Despite the risks associated with this process, granting AI a shared vocabulary and practices from cognitive science is not merely a concession but a necessary practice for research in the field of NLMs. Some examples of successful collaboration are highlighted in Angius *et al.* (2024), where it is demonstrated how a co-simulative method can help clarify the functioning and characteristics of both natural and artificial systems, using a classic cognitive science approach such as the synthetic method (Newell and Simon, 1972). One of the examples cited in the work involves biorobotics (Grasso *et al.,* 2000), where the construction of a robot lobster has aided in understanding the dynamics of the animal's foraging behavior. Conversely, another aspect concerns NLMs, where similarities are found between the activation in language processing by the transformer and the activation patterns in certain brain areas responsible for language (Caucheteux *et al*., 2023; Kumar *et al*., 2023).

A constructive anthropomorphism implies a full awareness of the different nature of human subjects subjected to cognitive investigations and of the NLMs to which they seek to apply the same. It is an awareness that is growing, with recent studies taking responsibility for it. For example, Löhn *et al*. (2024) stress how assesment procedures for psychological tests usually go through standard validation processes over the years; similar standardization processes are lacking when evaluating NLMs tests. In addition to the adequacy of the types of tests, it is important to consider the possible difference at the cognitive level, due to the different nature of the two entities. For example, since NLMs are trained on extremely vast corpora, they exhibit a cognitive style that reflects their advantage in information availability compared to a human.

Neglecting to allow a certain degree of "proper" anthropomorphism in AI studies would result in missing the opportunity for NLMs research to leverage the descriptive power of other fields–a practice that is helping to clarify and better understand the artificial phenomena studied within transformer architectures by relating them to studies and data obtained on humans. Finally, the sharing of a common scientific vocabulary, both in terms of description and practices, can only benefit both fields, enhancing the descriptive power and understanding of the functioning of both the human and algorithmic black boxes.

References

Alshawi, H. (1990), "Resolving quasi logical forms", in *Computational Linguistics*, 16:133–144.

Angius, N., Perconti, P., Acciai, A., & Plebe, A. (2024), *Co-simulations of brain language processing using neural language models*, in *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 4596–4603.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023), "Out of one, many: Using Language Models to simulate human samples", in *Political Analysis*, 31:337–351.

Bahdanau, D., Cho, K., & Bengio, Y. (2016), "Neural machine translation by jointly learning to align and translate", in *International Conference on Learning Representations*.

Banerjee, S. & Pedersen, T. (2003), " Extended gloss overlaps as a measure of semantic relatedness", in *International Joint Conferences on Artificial Intelligence*, pages 805–810.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021), "On the dangers of stochastic parrots: Can language models be too big?" , in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ACM, pages 610–623.

Bender, E. M. & Koller, A. (2020), "Climbing towards NLU: On meaning, form, and understanding in the age of data", in *58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Somerset (NJ). Association for Computational Linguistics.

Binz, M. & Schulz, E. (2023), "Using cognitive psychology to under- stand GPT-3", in *Proceedings of the Natural Academy of Science USA*, 120: e2218523120.

Bishop, J. M. (2021), "Artificial intelligence is stupid and causal reasoning will not fix it.", in *Frontiers in Psychology*, 11:513474.

Brunet-Gouet, E., Vidal, N., & Roux, P. (2023), "Can a conversational agent pass theory-of-mind tasks? a case study of ChatGPT with the hinting, false beliefs, and strange stories paradigms", in *Zenodo*, DOI 10.5281/zen- odo.8009748.

Bruni, D., Perconti, P., & Plebe, A. (2018), "Anti-anthropomorphism and its limits", in *Frontiers in Psychology*, 9:2205.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023), "Sparks of artificial general intelligence: Early experiments with GPT-4", in *arXiv*, abs/2303.12712.

Caron, G. & Srivastava, S. (2022), "Identifying and manipulating the personality traits of Language Models", in *arXiv*, abs/2212.10276.

Caucheteux, C., Gramfort, A., & King, J. (2023), "Evidence of a predictive coding hierarchy in the human brain listening to speech", in *Nature Human Behaviour*, 7:430–441.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Valter, D., Narang, S., Mishra, G., Yu, A. W., Zhao, V., Huang, Y., Dai, A. M., Yu, H., Petrov, S., hsin Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., & Wei, J. (2022), "Scaling instruction-finetuned language models", *arXiv*, abs/2210.11416.

Clippinger, J. H. (1977). *Meaning and Discourse – A Computer Model of Psychoanalytic Speech and Cognition*, Baltimore, Johns Hopkins University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", in *Proceedings North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics.l, pages 4171–4186.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985), "The satisfaction with life scale", in *Journal of Personality Assessment*, 49:71–75.

Eysenck, M. W. & Eysenck, C. (2022). *AI vs Humans*, Abingdon (UK); New York, Routledge.

Grasso, F. W., Consi, T. R., Mountain, D. C., & Atema, J. (2000), "Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges", in *Robotics and Autonomous Systems*, 30:115–131.

Groenendijk, J. & Stokhof, M. (1991), "Dynamic predicate logic", in *Linguistics and Philosophy*, 14:39–100.

Hagendorff, T. (2023), "Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods", in *arXiv*, abs/2303.13988.

Hagendorff, T., Fabi, S., & Kosinski, M. (2022). Machine intuition: Uncovering human-like intuitive decision-making in GPT-3.5. *arXiv*, abs/2212.05206.

Han, S. J., Ransom, K., Perfors, A., & Kemp, C. (2024), "Inductive reasoning in humans and large language models", in *Cognitive Systems Research*, 83:101155.

Haugeland, J. (ed. by) (1991), *Mind Design II*, Cambridge (MA), MIT Press, Second edition, 1997.

Heyes, C. & Frith, C. (2010), "The cultural evolution of mind reading", in *Science*, 344:1243091.

Hinton, G. & Zemel, R. S. (1994), "Autoencoders, minimum description length and Helmholtz free energy", in *Advances in Neural Information Processing Systems*, pages 3–10.

Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2022), "MPI: Evaluating and inducing personality in pre-trained language models", in *arXiv*, abs/2206.07550.

Joshi, A. K. & Schabes, Y. (1991), "Tree-adjoining grammars and lexicalized grammars. Technical Report MS-CIS-91-22", University of Pennsylvania Department of Computer and Information Science.

Karra, S. R., Nguyen, S., & Tulabandhula, T. (2022), "Estimating the personality of white-box language models", in *arXiv*, abs/2204.12000.

Karttunen, L., Koskenniemi, K., & Kaplan, R. M. (1987), "A compiler for two-level phonological rules", in Mary Dalrymple, R. K., Karttunen, L.,

Koskenniemi, K., Shaio, S., and Wescoat, M. (eds), *Tools for morphological analysis*, Center for the Study of Language and Information, Palo Alto (CA), Stanford University, pages 1–61.
Kay, M. (1984), "Functional unification grammar: A formalism for machine translation", in *Proceedings of the Twenty-Second Meeting of the Associa- tion for Computational Linguistics*, Association for Computational Linguistics, pages 75–78.
Kosinski, M. (2023), "Theory of mind may have spontaneously emerged in large language models", in *arXiv*, abs/2302.02083.
Koskenniemi, K. (1983), *Two-level morphology: a general computational model for word-form recognition and production*, PhD thesis, University of Helsinki.
Krause, S. & Stolzenburg, F. (2024), "From data to commonsense reasoning: The use of large language models for explainable AI", in *arXiv*, abs/2407.03778.
Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., & Nastase, S. A. (2023), "Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model", in *bioRxiv*, DOI: 10.1101/2022.06.08.495348.
Landgrebe, J., & Smith, B. (2019), "There is no artificial general intelligence", in *arXiv*, abs/1906.05833.
Lappin, S., & Leass, H. J. (1994), "An algorithm for pronominal anaphora resolution", in *Computational Linguistics*, 20:535–561.
Larson, E. J. (2021), *The myth of Artificial Intelligence: why computers can't think the way we do*, Cambridge (MA), Harvard University Press.
Lascarides, A., & Asher, N. (2007), " Segmented discourse representation theory: Dynamic semantics with discourse structure", in Bunt, H. and Muskens, R. (eds), *Computing Meaning*, Berlin Springer-Verlag, pages 87–124.
Leib, R. (2023). *Exoanthropology – Dialogues with AI*. Punctum Books, Brooklyn (NY).
Lesk, M. E., & Schmidt, E. (1975). Lex – a lexical analyzer generator. Technical Report 39, Bell Laboratories, Murray Hill (NJ).
Levin, E., Pieraccini, R., & Eckert, W. (1997), "A stochastic model of computer-human interaction for learning dialogue strategies", in *Eurospeech*, volume 97, pages 1883–1886.
Li, X., Li, Y., Liu, L., Bing, L., & Joty, S. (2022), "Is GPT-3 a psychopath? evaluating large language models from a psychological perspective", in *arXiv*, abs/2212.10529.

Löhn, L., Kiehne, N., Ljapunov, A., & Balke, W.-T. (2024), "Is machine psychology here? on requirements for using human psychological tests on large language models", in *International Natural Language Generation Conference*, Association for Computational Linguistics, pages 230–242.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013), "Distributed representations of words and phrases and their compositionality", in *Advances in Neural Information Processing Systems*, pages 3111–3119.

Nagel, E. (1961), *The structure of science*, Hackett, Indianapolis (IN).

Newell, A. & Simon, H. A. (1972), *Human problem solving*, Englewood Cliffs (NJ), Prentice Hall.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., & *et al.* (2022), "Training language models to follow instructions with human feedback", in *Advances in Neural Information Processing Systems*, pages 27730–27744.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023), "AI deception: A survey of examples, risks, and potential solutions", in *arXiv*, abs/2308.14752.

Perconti, P. & Plebe, A. (2023), "Do machines really understand meaning? (again)", in *Journal of Artificial Intelligence and Consciousness*, 10:181–206.

Perner, J., Leekam, S. R., & Wimmer, H. (1987), "Three-year-olds' difficulty with false belief: The case for a conceptual deficit", in *British Journal of Developmental Psychology*, 5:125–137.

Plebe, A. & Perconti, P. (2022), *The Future of the Artificial Mind*, Boca Raton, CRC Pres.

Pollard, C. & Sag, I. (1994), *Head-driven phrase structure grammar*, Chicago (IL), University of Chicago Press.

Rao, H., Leung, C., & Miao, C. (2022), "Can ChatGPT assess human personalities? a general evaluation framework", in *Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986), "Learning representations by back-propagating errors", in *Nature*, 323:533–536.

Rumelhart, D. E. & McClelland, J. L. (1986a), "On learning the past tenses of English verbs", in Rumelhart & McClelland (1986b), pages 216–271.

Rumelhart, D. E. & McClelland, J. L. (eds) (1986b), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge (MA), MIT Press.

Shardlow, M. & Przybyła, P. (2022), "Deanthropomorphising NLP: Can a language model be conscious?", in *arXiv*, abs/2211.11483.

Smith, G. (ed. by) (2018), *The AI Delusion*, Oxford (UK), Oxford University Press.

Tatar, D., Serban, G., Mihis, A. D., & Mihalcea, R. (2009), "Textual entailment as a directional relation", in *Journal of Research and Practice in Information Technology*, 41:53–64.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2023), "Do large language models know what humans know?", in *Cognitive Science*, 47: e13309.

Ullman, T. D. (2023), "Large language models fail on trivial alterations to theory-of-mind tasks", in *arXiv*, abs/2302.08399.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017), "Attention is all you need", in *Advances in Neural Information Processing Systems*, pages 6000–6010.

Vieira, R. & Poesio, M. (2000), "Processing definite descriptions in corpora", in Botley, S. P. & McEnery, T. (eds), *Corpus-based and Computational Approaches to Discourse Anaphora*, Amsterdam, John Benjamins, pages 189–212.

Webb, T., Holyoak, K. J., & Lu, H. (2023), "Emergent analogical reasoning in large language models", in *Nature Human Behaviour*, 7:1526–1541.

Weizenbaum, J. (1966), "Eliza – a computer program for the study of natural language communication between man and machine", in *Communications of the Association for Computing Machinery*, 9:36–45.

Williams, J. D., Poupart, P., & Young, S. (2005), "Factored partially observable Markov decision processes for dialogue management", in *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, AAAI press, pages 76–82.

Wimmer, H. & Perner, J. (1983), "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception", in *Cognition*, 13:103–128.

Wolfram, S. (2023), *What Is ChatGPT Doing ...and Why Does It Work*, Champaign (IL), Wolfram Media.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023), "The dawn of LMMs: Preliminary explorations with GPT-4V(ision)", in *arXiv*, abs/2309.17421.

Zarenchnak, M. & Brown, A. (1961), "Current research at Georgetown university", in Edmundson, H. (ed. by), *Proceedings of the National Symposium on Machine Translation*, Prentice Hall, Englewood Cliffs (NJ), pages 63–87.

Zhang, J., Hedden, T., & Chia, A. (2012), "Perspective-taking and depth of theory-of-mind reasoning in sequential-move games", in *Cognitive Science*, 36:560–573.

Bayesian “No Miracle Argument” and the Priors of Truth

Mario Alai
University of Urbino
Académie Internationale de Philosophie des Sciences
mario.alai@uniurb.it

1. Introduction

The “no miracle argument” (NMA) is generally considered the “ultimate” argument for scientific realism (Musgrave 1988). However, its roughest formulations (e.g.: “The success of science would be a miracle unless scientific theories were true”; “the only non-miraculous explanation of the success of science is scientific realism”) are as vague as open to many criticisms. Therefore, realists have already introduced several refinements to fend the objections which in fact have been raised. Some antirealists even concluded that this attempt to immunize it from objections is an endless and hopeless process, but I maintain that this is not the case, since there is at least one effective and undefeated form of the NMA. To show this, I begin by briefly discussing four refinements that have already been introduced (§ 2). Next, I introduce the “base-rate fallacy” objection, which until now has not received sufficient attention from realists (§ 3). Before answering it, I need to present yet another objection and the further refinement it calls for (§ 4). In § 5 I argue that the truth-conduciveness of the scientific method warrants a non-zero prior probability to hypotheses, and in § 6 I explain that for this reason the NMA avoids the base-rate fallacy. Finally, in § 7, I briefly reply to some further objections.

Mario Alai, “Bayesian ‘No Miracle Argument’ and the Priors of Truth”, in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 31-51

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

2. Four initial refinements of the NMA

Refinement (1)

That science in general is successful is evident but also quite vague, hence trying to explain its success is a “shaky game”. Therefore, although arguing for realism from the general success of science might be possible, like many others I prefer to discuss an argument from the success of particular theories. In fact, not all theories are successful, after all. As we shall see, however, even in considering a successful theory, we need to focus more specifically on some of its hypotheses. Still, as it turns out, such an argument needs to be supplemented by an account of the truth-conduciveness of scientific method in general (§ 5). Moreover, saying that many particular theories are successful is saying that science in general is often successful. Therefore, claiming that realism is the only plausible explanation of the success of theories is also claiming that it is the only explanation of the success of science in a rather straightforward sense. Thus, an initial refinement of the NMA is:

(i) The only (non-miraculous) explanation why a theory has success is that it is true.
(ii) Theory T has success.

Therefore

(iii) T is true (save miracles).

Notice, the strength of (i) is such that, assuming that every event has (at least) one explanation, this argument is no longer abductive, but deductive[1].

Refinement (2)

As it is, premise (i) is false, because various types of success can be explained without assuming that a theory is true. Success in accommodating previously known phenomena is explainable by the skill and patience of theoreticians. The prediction of phenomena similar to the already known ones can be explained by analogical or inductive extrapolation. What needs to be explained, instead, is *novel* success, i.e., the prediction of phenomena that were previously unknown, or at least neither used in constructing the theory, nor similar to those used (Alai 2104a, §§ 3.3, 3.4).

[1] See also Golemon & Graber (2023).

Refinement (3)

Even a completely false theory can entail true consequences. For instance, a theory comprising the claims 'Paris is the capital of Italy, and the Colosseum is in Paris' correctly predicts that the Colosseum is in the capital of Italy. Thus, it is no wonder that false theories get right certain "easy" (i.e., probable) predictions. For instance, if an astrological theory predicts that the next number on the roulette will be even, the chance that it turns out right is about 0,5. Hence, any predictive success of this kind can be explained as due to moderate luck, without assuming that the theory is true.

On the contrary, it is extremely unlikely that a false theory gets right a very improbable prediction[2]. Based on Newton's theory and the irregularities of Uranus' orbit, Leverrier predicted the existence of a new planet (later called 'Neptune') and its position with an error of less than 1°. Since there are 360° on the horizon and 360° on the altitude, the probability that by chance a false theory predicted the right position with an approximation of ±1° was 2/360=1/180 on each axis, and the joint probability was 1/180·180 = 0.00003. Other predictions are even less probable: the prediction of the magnetic moment of the electron made by quantum electrodynamics was accurate to the 9th decimal, so its probability was 0.000000001 (Wright 2002: 143–144). Thus, it is only the success of novel and improbable predictions that must be explained by the truth of the theory.

Refinement (4)

Larry Laudan (1981) pointed out that in the past many false theories made nonetheless true and improbable novel predictions, considering this as a *reductio* of the claim that novel predictions warranted the truth of theories. Newton's gravitational theory is a case in point, since it is false in spite of its just mentioned striking prediction. Deployment realists, especially Kitcher (1993) and Psillos (1999), replied that in those cases only certain hypotheses of the theory had been deployed in deriving the successful prediction, hence the only explanation for its success is that at least those hypotheses were true, while the rest of the theory could well be false.

Timothy Lyons rejoindered (2002, 2006) by listing many individual hypotheses now considered false that had been deployed in novel predictions. However, it has been remarked[3] that often the hypotheses from which a

[2] Alai (2014a): § 3.2, (2014b): §§ 4, 5.
[3] Psillos (1999), Alai (2014a): 307; (2014b): 268-269, § 7; (2021).

prediction is derived are redundant, i.e., not *essential* to its derivation. For instance, suppose that a neo-pagan theology claims that

(H) When the barometer is low, Zeus sees to it that it rains.

If the theorist observes that the barometer is low, she can predict that it will rain and turn out right. Thus, H is deployed in a successful prediction, yet it is false. However, only a part of H is essential to that prediction, viz.,

(H') When the barometer is low, it rains,

and sure enough, it is true. A real example is again the prediction of the existence of Neptune. Like many others, it was derived from Newton's false hypothesis that

(N) Bodies are moved by a gravitation force proportional to their masses and inversely proportional to the square of their distance, and space is flat.

(N) is false, because there is no gravitation force and space is curved, but (N) was not essential, only its true part played an actual role in the derivation:

(N') 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 approximates Newton's law[4].

So, a hypothesis H deployed in predicting the novel phenomenon NP is not essential to that prediction if it entails a weaker hypothesis H' which in turn entails NP. Even H' may fail to be essential, as it may entail a weaker H'' still entailing NP. Only the weakest hypothesis still entailing NP is deployed essentially in the prediction. Any hypothesis H deployed unessentially may be false, but by definition it entails a hypothesis H^e which is essential to predicting NP, hence is certainly true. H^e is part of the content of H, hence of the content of the theory T to which H belongs. Therefore, except for miracles, a hypothesis from which a risky novel prediction has been derived is at least *partly* true i.e., it has some true content, and so is the theory to which it belongs[5]. Notice, even the weakest hypothesis deployed to predict NP must

[4] For various other examples see Alai, (2014b): 282-286; (2021): 192-198; (2024): 256.

[5] It may be difficult to ascertain whether a hypothesis H, from which a novel prediction NP was derived, was essential to that prediction or not, hence whether we can be assured that it is (completely) true or not (Alai 2021: 199-204). However, we know that if it is not true, it is at least partly true.

still be theoretical, not empirical, because mere empirical claims could not entail novel and improbable predictions. Besides, empirical claims can be checked by observation, hence they don't need top-down confirmation by consequences.

A fourth refined formulation of the NMA is therefore:

(i') The only (non-miraculous) explanation why a theory T predicted a novel (i.e., not used or similar to those used) and improbable phenomenon is that the hypothesis H of T deployed in the prediction is true (if it was essential to that prediction), or partly true (if it was not essential).

(ii') T predicted a novel and improbable phenomenon NP.

Therefore,

(iii') save for miraculous coincidences, the hypothesis H of T deployed in the prediction of NP (hence, T itself) is at least partly true.

In what follows, therefore, the truth of theories and hypotheses will always be understood as *at least partial*. At any rate, objections have been raised even against this formulation of the NMA, as we shall see now.

3. Bayes' theorem and the "base-rate fallacy" objection

Objection (1)

Saying that a hypothesis H entails the prediction of the novel phenomenon NP is saying that NP becomes certain[6] if we assume that H is true. Therefore, the fourth refined formulation of the NMA can be expressed also in this way: since NP has a very low prior probability (e.g., 0.0003), but it becomes certain if one assumes H, it is highly probable that H is true. It has been objected, however, that this commits the *base-rate fallacy*: the probability of H given NP—p(H|NP)—cannot be computed only from the prior probability of NP—p(NP)—and the conditional probability of NP given H—p(NP|H)—, which in this case is 1. As shown by Bayes' theorem, p(H|NP) depends also on the prior probability that H is true—p(H)—:

[6] Or gets a certain specifiable probability if the prediction is not deterministic but probabilistic.

$$\text{(Bayes theorem) } p(H|NP) = \frac{p(NP|H)\ \cdot p(H)}{[P(NP|H) \cdot p(H)] + [p(NP|\neg H)\ \cdot p(\neg H)]}$$

Now, many antirealists argue that, due to the empirical underdetermination of theories, there are infinitely many false hypotheses and only a true one compatible with all the empirical data[7]. Moreover, in their view, the only epistemic support for hypotheses comes from their relationship with the data, since citing support from other hypotheses would be question-begging, and considerations like simplicity, elegance, etc. are only pragmatically relevant. Therefore, they claim, there are infinitely possible and equally probable alternative hypotheses, hence the prior probability of each one tends to zero ($p(H) = 1/\infty = 0^+$). Consequently, also its conditional probability tends to zero ($p(H|NP) = 0^+$)[8]:

$$p(H|NP) = \frac{1\ \cdot 0^+}{(1 \cdot 0^+) + (0{,}00003\ \cdot 1^-)} = 0^+$$

This reasoning, however, has the paradoxical consequence that no hypothesis can ever be confirmed by any prediction or any empirical evidence whatsoever. This conclusion, of course, crucially depends on assuming that $p(H) = 0^+$. In fact, as we shall see, if p(H) is even slightly greater than $1/\infty$ and NP is improbable, p(H|NP) increases dramatically. Moreover, if p(H) is updated in the light of a few more predictions NP', NP'', etc., by taking as the new prior probability of H first P(H|NP), then P(H|NP'), etc., it soon converges to 1. Thus, we must ask: does really p(H) equal 0^+? To answer this question, we must first consider another possible objection to our NMA.

---

[7] An anonymous referee rightly pointed out that scientists only consider finitely many serious hypotheses at one time. However, Stanford (2006) argued that scientists systematically fail to consider many relevant alternatives, including some which are more probably true. At this stage, therefore, we cannot a priori rule out any logically possible alternative hypothesis without begging the question against antirealists. Showing how a prior probability which initially is equally distributed over infinite hypotheses eventually can collapse almost completely on just a few ones is precise the task of the ensuing argument.

[8] Howson (2000), Magnus & Callender (2003), Dieks (2024): 113, Morganti (2024): 127.

4. How hypotheses entailing novel predictions can be found

Objection (2)

The NMA claims that the only non-miraculous explanation of T's success in predicting NP is the truth of the hypothesis H deployed in the prediction. However, as noticed by Roger White (2003: 659-663), if by 'T', 'H' and 'NP' we rigidly refer to a certain theory, a certain hypothesis and a certain phenomenon, the prediction of NP by T is trivially explained simply by the logical fact that T, and more particularly H (together with the appropriate background assumptions) entail NP. There is no need to assume that H is true. After all, as noticed earlier, a completely false hypothesis can entail true consequences (Alai 2014a: 299).

One has the immediate impression that this objection misses something, but it cannot be rejected if one sticks literally to the quick formulation of the NMA according to which only the truth of H explains T's prediction of NP. Therefore, in order to resist this objection and to bring out what it misses, we need to formulate the argument more explicitly by a further refinement:

Refinement (5)

Consider this: all possible consistent hypotheses entail a tautology, no one entails a contradiction, and in general, the less probable a prediction is, the fewer hypotheses entail it. Here, as before, we are speaking of probability in a purely a priori sense, as the inverse of the informative content of a hypothesis. In this sense, saying that the probability of NP is (e.g.) 0.000000001 is saying that, by gross approximation, NP is entailed by about 1 hypothesis out of 1,000,000,000 possible ones, and by a negligible proportion even of the possible hypotheses compatible with the already known data. If the probability of NP is 0.00003, it will be entailed by approximately 3 out of 100.000 possible hypotheses. Therefore, what must be explained is

(Q) How have scientists been able to find a theory T which included H, one of the extremely rare hypotheses entailing NP?[9]

Surely, they didn't construct it in order to entail NP, because NP was novel, nor they picked it randomly, because the chance to find it was in the order of 0.00003 for the position of Neptune, and 0.000000001 for the magnetic moment of the electron. In other words, it is almost certain (e.g., there is a probability of something like 1 − 0.00003, or 1 − 0.000000001) that

[9] Alai (2014a): 299, White (2003): 659–663.

T and H were not chosen randomly, but *by an effective procedure*. Granted, if two, or three, or … *n* different hypotheses are tried by scientists, the probability that one of them entails NP becomes 2 or 3 or … *n* times higher[10], but still remaining very low, and seldom more than a few attempts are made. For instance, if astronomers had tried 10 different models of Neptune, the probability of predicting its position would have been 0.00003·10 = 0.0003.

The procedure by which scientists conceive theories is *scientific method* (SM). Moreover, the rate of theories chosen by scientists which yield novel predictions is fairly high, quite higher than the probability of getting such theories by picking them randomly. For instance, the frequency of theories entailing novel predictions as improbable as Neptune's position is substantially higher than 0.0003, and the frequency of theories entailing novel predictions as improbable as the magnetic moment of the electron is substantially higher than 0.00000001. So, SM is an effective method to find theories yielding novel predictions. Moreover, as we shall soon see, it can be so effective simply because it is reliable in finding true theories.

Does this mean that the context of discovery (how theories are generated) matters to the context of justification (i.e., to how their plausibility is evaluated)?[11] In certain senses yes, but in other equally relevant senses, no. I will explain this after fully developing my proposal, at the end of § 6.

5. The truth-conduciveness of scientific method

Why is SM so effective in generating theories which entail novel predictions? Because true theories entail true consequences, and if they are strong enough, they also entail true, novel, and very informative (i.e., improbable) novel consequences, like NP. Therefore, if we assume that fairly frequently science produces true and sufficiently strong theories, we can explain why rather frequently it also produces startling novel predictions. Granted, science also produces many (completely) false hypotheses, but it would be a miraculous coincidence if one of those happened to entail a risky novel prediction. That is, SM leads to novel predictions not simply because it is, generically, reliable, but because—and to the extent that—in certain occasions it actually succeeds in generating true hypotheses: it is practically certain that the hypotheses from which risky novel predictions are derived are true.

---

[10] See Dawid, Hartmann (2018): § 8.

[11] I how this question to an anonymous referee.

Scientific antirealists, however, deny that SM is so reliable in tracking the truth: why are we entitled to assume that it is? Because theories and hypotheses are not just arbitrarily imagined and then tested top-down for success. Rather, SM provides a reliable heuristic through which they gradually emerge bottom-up respecting empirical constraints and thus earning a certain degree of confirmation since the very beginning[12]. The data on which they based are not thoroughly theory-relative, because the reliability of observational instruments, even the most sophisticated, is ultimately warranted by direct observation through a *recursive empirical foundation* (REF) process. Whenever hypotheses are introduced that cannot be inferred inductively from observation, they are not fully accepted until they are strongly confirmed by highly reliable top-down controls.

Very schematically, scientific discovery proceeds through the following steps:

(1) from direct observation by induction we infer empirical generalizations about observable but not yet observed phenomena. If the correct inductive methods are used, these conclusions are highly confirmed, moreover they can be checked by direct observation.

(2) By measurement of direct observable quantities with the aid of some elementary mathematics we establish claims about entities which are unobservable only because they are smaller than the observable ones. For instance, this is how Perrin measured the size of molecules and Millikan the charge of electrons.

(3) By abduction, analogy and inference to common causes we infer from observed phenomena to non-observed ones. When we use these inference patterns to predict unobserved but observable phenomena, subsequently we can check by direct observation whether their conclusions obtain or not. In this way we realize that, although in general they are less reliable than induction (let alone deduction), they are far more reliable than random guessing. That is, when their premises are true, their conclusions also prove true with a probability $0<<p<<1$ which in general is distinctly lower than 1 and higher than 0. This probability differs sensibly for each specific inference, depending on a host of factors; hence, figuring it more precisely is difficult in the particular case, and impossible in general.

Unlike induction, however, abduction, analogy and inference to common causes can be used also as *theoretical* inferences, i.e. inferences from observations to theoretical claims about unobservable entities or phenomena.

---

[12] A bottom-up strategy in the confirmation of hypotheses and in the defence of scientific realism is strongly advocated by Michel Ghins (2024).

For instance, Vitruvius arrived at a wave theory of sound by analogy to water waves[13]. In such cases their conclusions cannot be checked by direct observation, but they have at least the probability $0<<p<<1$ provided by the general reliability of these inference patterns.

(4) There are instruments by which we can observe both entities which are observable also by the naked senses and others which are not. By direct sensorial observation we realize that these instruments are reliable when used for sensorially observable entities, and by induction we trust that they are equally reliable for the directly unobservable entities.

For example, in Venice Galileo demonstrated the reliability of his astronomical telescope by asking bystanders to observe through it the city of Chioggia across the lagoon; thus, they realized that what they saw was exactly what they could see by the naked eye at a closer distance. Analogously, Van Leeuwenhoek, a cloth merchant, originally used his rudimentary optical microscope to gain enlarged images of his fabrics, hence he knew he could trust it when it showed him the first bacteria ever observed.

Philip Kitcher (2001) called this strategy “Galilean”, and we can call “Galilean” both the instruments checked by the Galilean strategy, and the observations made by those instruments. Galilean validation is recursive: for instance, after establishing the reliability of optical microscopes by direct observation, we can establish that of electronic microscopes by using them to observe tiny objects which can also be seen through optical microscopes, and so on.

(5) Through Galilean observation we directly (bottom-up) discover certain unobservable truths. For example, van Leeuwenhoek observing through his microscope discovered that there exist bacteria.

(6) Through Galilean observation we can test at least some of the theoretical claims introduced at step (3) and some of them are *confirmed*, while others are discarded. For instance, Galileo’s claim that celestial bodies have the same nature as the Earth—e.g., have mountains—was confirmed by observing through the telescope.

(7) Starting from Galilean observation (4), deeper theoretical claims can be derived by further theoretical inferences (abductions or inferences by analogy).

(8) The conclusions achieved by steps (1), (2), (5) and (6) are *highly probable* and *firmly believed*, since they have a strong direct or indirect empirical (bottom-up) support. Instead, the claims advanced at step (3) that

---

[13] Vitruvius (1960): 138-139; Holland, Holyoak, Nisbett & Thagard (1986), Ch. 10; Gentner, Holyoak & Kokinov (2001): 7.

cannot be tested as at step (6) and those advanced at step (7) are introduced by theoretical inferences, hence they are much less confirmed, i.e., they have only a probability 0<<p<<1 corresponding to the general reliability of those inferences[14]. Their probability is augmented by two factors: (i) the indirect support these conclusions may receive from their role in the context of the whole theory, as shown by Hempel's suspended net model[15]; (ii) the requirement to be consistent with all of our most probable and firmly believed claims[16]. Even so, while their probability is clearly higher than zero, it is still far from 1. These claims, therefore, are not firmly believed, yet, but held only as *hypotheses*.

(9) At least some of these hypotheses, however, can be confirmed top-down in two ways:

(i) first, if they entail novel improbable predictions borne out by direct (steps 1, 2) or indirect (step 4) observation. This is the most common formulation of the NMA, since when the prediction is highly improbable, it is also highly improbable that the hypothesis is false (it would be a miracle if it were false). For example, Fresnel's wave theory was confirmed by the prediction of white spot, while Einstein's relativity theory was confirmed by the predictions of the gravitational deflection of light rays, by the retard of clocks in motion, etc.

(ii) The hypotheses advanced at steps (3) and (7) can be confirmed also by "consilience", i.e., if they coincide with the conclusions reached from independent premises and/or by independent methods, or if they are confirmed by instruments, even non-Galilean but based on independent firmly believed theories[17]. An example is Perrin's (1913) measurement of Avogadro's number by independent methods. It is arguable that certain instances of consilience would be "miraculous" coincidences if the

---

[14] To be precise, if 0<<p<<1 is the probability that the conclusion of a theoretical inference is true when its premises are true, the probability of the conclusions of the inferences (3) and (7) is only slightly less, since their premises are highly probable.

[15] "A scientific theory might be linked to a complex spatial network. Its terms are represented by the knots, while the threads connecting the latter correspond, in part, to the definitions and, in part, to the fundamental and derivative hypotheses included in the theory. The whole system floats, as it were, above the plane of observation and is anchored to it by the rules of correspondence. These might be viewed as strings which link certain points [of the network] with specific places in the plane of observation" (Hempel 1952: 36).

[16] In practice, when a new hypothesis is otherwise very promising, certain contradictions with previously firmly believed hypotheses are tolerated, but they are considered a problem for it.

[17] See Kosso (1992): Ch. IX.

hypotheses were false, just like novel predictions, hence they may confirm to the same degree[18].

It is noticeable that:

• The claims advanced at steps (1), (2), (4), (5) and (6) are highly probable and firmly believed because they are ultimately based only on direct observation plus elementary computations and/or induction. This should be acknowledged even by strict empiricists. One could radically doubt those claims only by embracing Humean or Cartesian scepticism (i.e., by denying in principle the reliability of induction or of perception).

• The hypotheses confirmed by Galilean observation at step (6) and by novel predictions and consilience at step (9) become highly probable or even practically certain, hence are firmly believed.

• Testing hypotheses by Galilean observation at step (6) also allows us to discard many prima facie plausible but false alternatives, thus approaching the truth also by elimination.

• At this stage, therefore, only few claims are still held *merely* as hypotheses, just on the basis of theoretical inferences like abduction or analogy, without further tests.

(10) Through the strongly confirmed knowledge provided by steps (1), (2), (4), (5), (6) and (9) we build and validate *new more sophisticated instruments*. They are Galilean in a wider sense, since they are validated based on highly probable claims.

(11) Based on the achievements of steps (1), (2), (4), (5), (6), (9) and (10), by induction, measurement, or direct discovery we reach new strongly confirmed claims.

(12) Based on the achievements of steps (1), (2), (4), (5), (6), (9) and (10), by theoretical inferences we advance still deeper hypotheses, which at this stage are held only hypothetically, with a probability $0<<p<<1$, as explained at (8) above.

(13) The hypotheses advanced at step (12) are tested by Galilean observation, and/or by consistency with previously firmly believed hypotheses, and/or by novel predictions and/or by consilience of independent methods. If confirmed, they become highly probable and firmly believed.

[18] Alai (2014a). As explained in § 2, in general these top-down tests can confirm to a high degree that hypotheses are *at least partly* true. Thus, even the hypotheses confirmed by them may subsequently be substituted by others with a larger true content or a smaller false content. This is why certain past theories or hypotheses that were confirmed in these ways and firmly believed are no longer accepted today: our current theories preserve their true content but have dropped part of their false content.

(14) Based on knowledge reached at steps (12) and (13) we build and validate yet new more sophisticated instruments, and so on, recursively.

We may call this process the "*Recursive Empirical Foundation"* (REF) of theories. Summing up, at each time the majority of our theoretical claims are firmly believed and actually very probable.

Scientists[19] often state that they don't *really* believe in a theory until they "see" the particles it postulates or the effects it predicts. Of course, by 'seeing' they don't mean direct sensorial observation but either instrumental observation, or the application of the NMA to very specific hypotheses. For instance, take the detection of Higgs boson (predicted in 1964) by the LHC at Cern in 2012, or the detection of gravitational waves (predicted by Einstein in 1916) by LIGO e VIRGO interferometers in 2015. Detections like these actually consist in the exact verification of extremely precise novel predictions entailed by those hypotheses[20]. In any case, "observations" of this kind are so sensational because, even if occurring after the theory has been widely accepted for many decades, they are considered as its definitive proof. They mark the passage from more or less hypothetical acceptance to firm belief. They show that scientists (besides fully relying on the confirming power of novel predictions) trust that their apparatuses, complex and sophisticated as they are, are based on highly probable assumptions and ultimately warranted by direct observation through the REF. Unlike van Fraassen (2024), scientists don't believe that these "observations" are completely theory laden.

6. The right prior probability of hypotheses and the resulting conditional probability

We are now ready to go back to the NMA and to the base-rate objection. In order to face it, I asked: does the prior probability that a hypothesis H is true tend to zero (i.e., is $p(H) = 0^+$)? The foregoing discussion shows that the answer is negative, because hypotheses are not chosen randomly, but through the SM. SM ensures that the hypotheses which need confirmation by the NMA (steps (9) and (13) above) have a probability $0<<p<<1$ distinctly higher than zero. I noticed that this probability is difficult to figure in general, as it depends on each particular hypothesis and on how it was reached (step 3), but it is somewhat raised by consistency constraints (step 8).

[19] Among them my colleague, physicist Catia Grimani.

[20] My former student, physicist Giulia Callisesi, attracted my attention to this.

Those constraints force scientists to weed out most logically conceivable hypotheses (steps 6, 12), including many empirically viable ones. Therefore, on any particular subject only relatively few hypotheses are actually considered over time, so that the frequency of the true hypotheses over those actually proposed is often quite high. For instance, concerning the structure of the solar system, only a handful of basic models were advanced: Ptolemy's, Copernicus', Brahe's, Riccioli's, besides Kepler's basically true model. On the structure of light only two basic hypotheses (corpuscular and undulatory) were considered before the currently accepted one. Concerning the structure of the atom, only five or six models have been proposed, etc. Even considering the variants of each hypothesis, the frequency of the true ones is still fairly high.

From this historical point of view, therefore, it might seem that the probability of a hypothesis prior to its confirmation by novel predictions or consilience might range approximately between 0.1 and 0.5. Admittedly, this would be too quick and oversimplified in various ways, and earlier I maintained only that the prior probability of a typical hypothesis is distinctly higher than zero. Therefore, for the sake of the argument, let's stick to a safe estimate, assuming for instance that it is 0.02. Thus, if we consider a hypothesis H which yielded a not extremely improbable novel prediction, like that of Neptune, we can compute its conditional probability p(H/NP) as follows:

$$p(H|NP) = \frac{P(NP|H)=1 \cdot p(H)=0.02}{[P(NP|H)=1 \cdot p(H)=0{,}02]+[p(NP|\neg H)=0.00003 \cdot p(\neg H)=0{,}98]} =$$

$$p(H|NP) = \frac{0.02}{[0{,}02]+[0.00003 \cdot 0{,}98]} =$$

$$p(H|NP) = \frac{0.02}{[0{,}02]+[0.0000294]} =$$

$$p(H|NP) = \frac{0.02}{[0.0200294]} = 0{,}9985321577$$

That is, an improbable novel prediction makes the (at least partial)[21] truth of the hypothesis deployed in it practically certain. Even if one complained that my 0,02 prior probability is too optimistic, things don't change radically. Suppose one is so pessimist to suggest that p(H) = 0.0001, i.e. that only one out of 10,000 hypotheses advanced by scientists is true. Even in this case p(H|NP) would come out as

---

[21] As explained in § 2.

$$p(H|NP) = \frac{0.0001}{0.0001+[0.00003 \cdot 0{,}9999]} = 0{,}7692485211$$

That would still be a significant confirmation, but if H produced also another independent and improbable prediction NP', we could update our assessment by using this value as the new prior probability of H, and the conditional probability p(H|NP') would become 0,999991001.

Therefore, a fully explicit formulation of the NMA, providing an answer to question (Q) of § 4, is approximately as follows:

(i") When scientists found a theory T predicting a novel (i.e., not used or similar to those used) and improbable phenomenon NP, the only (non-miraculous) explanation is that, thanks to the truth-conduciveness of the SM, they conceived a theory which included at least a true or partly true hypothesis H entailing NP.
(ii') T predicted a novel and improbable phenomenon NP.
Therefore,
(iii') save for miraculous coincidences, the hypothesis H of T deployed in the prediction of NP (hence, T itself) is at least partly true.

Therefore, an effective defence of scientific realism must acknowledge that the NMA is a necessary but not sufficient component of scientific practice. It is needed to raise the low prior probability 0<<p<<1 of hypotheses to a higher degree that warrants firm belief; in turn, however, it needs the bottom-up REF procedure to ensure that those prior probabilities are at least 0<<p<<1. This also seems a fair solution of the lengthy inductivism-deductivism debate in philosophy of science: induction and deduction, bottom-up and top-down inferences are both required and complementary.

In § 4 I asked whether the role I am attributing to SM in the confirmation of hypotheses means that the context of discovery (how hypotheses are generated) matters to the context of justification (i.e., to how their plausibility is evaluated). Now we can see that in certain important senses it does, but in other also relevant senses it does not.

Certain neopositivists and Popper held that the context of discovery is completely irrelevant to that of justification, for the degree of confirmation of a hypothesis depends exclusively on its logical relationship to the available data. Thus, whether a hypothesis was produced by orthodox research methods or chosen randomly, or discovered in dream (like Kekulé's benzene's ring) doesn't matter to its evaluation. This is right in the sense that, *in principle*, one might dream or *arbitrarily* imagine a hypothesis and then check it by observation, induction, abduction, or empirical control of its consequences:

in this way the context of discovery would be irrelevant to its confirmation or rejection.

In practice, however, hypotheses are not *arbitrarily* imagined but gradually developed from observation through the REF: universal empirical claims are conceived by observing particular instances and by generalizing, i.e., by the same inductive procedure through which they are confirmed. Granted, in the discovery of theoretical hypotheses a major role is played by imagination and subjective considerations which provide only very weak justification; therefore, hypotheses must be confirmed top-down, by comparing their empirical consequences with the data. However, the main claim of this paper is that such confirmation also depends on the prior probabilities, which are provided by the discovery process. Thus, discovery and justification proceed hand in hand.

There is a point, however, in stressing the preeminence of justification over discovery, since in the justification of theoretical hypotheses the top-down control by consequences has a far greater impact than the bottom-up confirmation provided by the discovery process: we have seen that prior probabilities are very uncertain and generally quite low, while successful novel predictions and consilience can raise those probabilities up to practical certainty.

On the other hand, when philosophers like Kuhn and Feyerabend stressed the relevance of the context of discovery, they understood it as including extra-scientific factors, like metaphysical presuppositions, traditional beliefs, or sociological drives. It followed that one and the same hypothesis could be considered highly confirmed by one community and not confirmed by another community with different metaphysical presuppositions, traditional beliefs presuppositions or sociological structures. Of course, nothing in my account supports the relevance of the context of discovery in this radical sense.

7. Further objections

Objection (3)

Dennis Dieks objected that any novel phenomenon NP is not entailed just by a true hypothesis H^T, but also by all the false hypotheses which are incompatible with H^T but "predictively similar" to it[22] or empirically

[22] I.e., license exactly the same predictions. He first objected this to me in April 2008 at the botanic garden of Utrecht University, where he had invited me to lunch.

adequate. More generally, given the empirical underdetermination of theories, in principle there are infinitely many false hypotheses which, besides being compatible with all the previously known phenomena, also entail NP. Therefore, if I come up with a hypothesis H^{my} which saves all the available data and also turns out to predict NP, it is much more probable that H^{my} is one of those false hypotheses, rather than the true one. Hence, novel success does not confirm.

The reply is that for any hypothesis which both saves the previously available data and happens to entail NP, there are infinitely many others which also save the previous evidence but fail to predict NP. Therefore, if hypotheses were conceived just by trying to save the known data, without any strategy for finding *true* hypotheses, the probability of finding one entailing NP would be proportional to the probability of NP. That is, in cases like the above-mentioned ones it would be practically impossible[23].

Objection (4)

Dieks also claimed that we can look for empirically adequate or predictively similar hypotheses not by random choice, but through a method, i.e., SM itself, just like we look for true hypotheses. Thus, we will find hypotheses entailing NP even more easily than by looking for true hypotheses (Dieks 2024: 116-117).

The response is that there is no method for finding sufficiently strong hypotheses which are empirically adequate, or predictively similar, *without* being also true (Alai 2014c: 57-61; 2024: 258): one can draw reliable empirical predictions either from true theoretical hypotheses, or by analogy and induction from observed phenomena. The latter strategy, however, allows one to predict only phenomena that are similar to the observed one, while novel predictions concern radically heterogeneous phenomena. SM is no exception: it leads to novel predictions (sometimes, not always) only in so far as it leads to the truth: if a hypothesis was conceived by the best possible scientific practice but happens to be false (as is quite possible), the chance that it is empirically adequate or predictively similar, or that it produces any novel predictions, is negligible. Summing up, SM could not be effective in finding hypotheses which yield novel predictions if it were not reliable in finding true hypotheses.

---

[23] See Alai (2012): footnote 6; (2014a): 299; (2014c): 50.

Objection (5)

Against steps (3) and (4) of the REF van Fraassen (1980) and many others have objected that instruments or inference patterns which have proven reliable for directly observable entities cannot be trusted for directly unobservable entities without begging the question.

The answer is that observability is not an intrinsic property of entities, it only depends on the specific properties of human sense organs, which have no causal influence on the physical relation between certain instruments and certain entities, or on the argumentative soundness of certain inference patterns. Typically, the only intrinsic difference between directly observable and unobservable entities is in size, but observation itself shows that in general size does not significantly affect the behaviour of entities. When it does, of course, this can also be recognized through the REF and taken into due account[24]. Without assuming the uniformity of nature (i.e., that similar things behave similarly in any respect R, except when they differ in ways causally affecting R) even elementary empirical beliefs could not be supported.

8. Conclusion

I have accounted for five successive refinements of the NMA, required to defend it from various objections. As a result, the argument can also be expressed as a syllogism, whose major premise is that hypotheses can be confirmed up to practical certainty by updating of their probabilities in light of their successful novel predictions: if a hypothesis H entails a very improbable prediction NP, it is very probably true.

It has been objected, however, that this reasoning commits the base-rate fallacy, since it overlooks the prior probabilities of H. According to Bayes' theorem, the conditional probability of H depends also on its prior probability, and if the latter tends to zero, even the former vanishes. Moreover, antirealists claim that the prior probabilities of hypotheses do tend to zero, because of the empirical underdetermination of theories.

I have replied that, on the contrary, they are distinctly higher than zero. This is because, while science is fallible, SM gradually and recursively extends empirical support from observative claims to theoretical hypotheses farther and farther from direct observability. Therefore, the NMA is correct even from a Bayesian point of view.

[24] Kitcher (2001): 174, 178; Alai (2010), § 3, p. 270.

Inductivist methodologists of science, like Mill and Reichenbach, believe in the bottom-up construction and validation of hypotheses, while deductivists, like Whewell and Popper, believe in top-down control and confirmation or elimination. A consequence of my discussion, however, is that these two kinds of procedures are both needed and complementary. In parallel, an effective defence of scientific realism must supplement the top-down strategy of the NMA with a full appreciation of the bottom-up support provided by SM to theoretical claims (and *vice versa*).

I also argued that what the NMA makes highly probable is not just empirical adequacy, but truth. More precisely, when a theory, or even a single hypothesis, entails a successful novel prediction, we can trust that it has at least some true part, viz., the part that played an *essential* role in the prediction. It is difficult, however, to know whether hypotheses are deployed essentially, and in general they are not. Therefore, we should expect that not only a theory but even a single hypothesis was redundant to its successful predictions, hence it probably includes also some falsities.

Acknowledgments

I thank Vincenzo Fano and two anonymous referees for many useful suggestions.

References

Alai, M. (2010), "Van Fraassen, Observability and Belief", in D'Agostino, M., Giorello, G., Laudisa, F., Pievani, T. & Sinigaglia, C. (eds.), *New Essays in Logic and Philosophy of Science*, SILFS vol. I, London 2010, College Publications, pp. 663-675.

Alai, M. (2012), "Levin and Ghins on the "No Miracle" Argument and Naturalism", in *European Journal for Philosophy of Science*, 2, n. 1, pp. 85-110.

Alai, M. (2014a), "Novel Predictions and the No Miracle Argument", in *Erkenntnis* 79, n. 2, pp. 297-326.

Alai, M. (2014b), "Defending Deployment Realism Against Alleged Counterexamples", in Bonino, G., Jesson, G., & Cumpa, J. (eds.), *Defending Realism. Ontological and Epistemological Investigations,* Boston-Berlin-Munich, De Gruyter, pp. 265-290.

Alai, M. (2014c), "Why Antirealists Can't Explain Success" in Bacchini, F., Caputo, S., &Dell'Utri, M. (eds.), *Metaphysics and Ontology Without*

Myths, Newcastle upon Tyne, Cambridge Scholars Publishing, pp. 48-66.

Alai, M. (2021), "The Historical Challenge to Realism and Essential Deployment", in Lyons, T., and Vickers P. (eds.), *Contemporary Scientific Realism. The Challenge from the History of Science*, New York, Oxford University Press, pp. 183-215.

Alai, M. (2024), "Comments and Replies to My Friends", in Angelucci, A., Fano, V., Ferretti, G., Galli, G., Graziani, P., & Tarozzi, G. (eds.) (2024), pp. 237-294.

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, Franco Angeli.

Dawid, R., & Hartmann, S. (2018), "The No Miracles Argument Without the Base-Rate Fallacy", in *Synthese*, n. 195, pp. 4063–4079.

Dieks, D. (2024), "The Probabilistic No-Miracle Argument", in Angelucci, Fano, Ferretti, Galli, Graziani, and Tarozzi (eds.) (2024), pp. 111-119.

Gentner, D., Holyoak, K.J., & Kokinov, B.N. (2001), "The Place of Analogy in Cognition", Introduction to: Gentner, Holyoak, & Kokinov (eds.), *The analogical mind: Perspectives from cognitive science*, Cambridge, MA, MIT Press.
https://www.academia.edu/119199808/The_place_of_analogy_in_cognition

Ghins, M. 2024, *Scientific Realism and Laws of Nature: A Metaphysics of Causal Powers*, Cham, Springer Verlag.

Golemon, L., & Graber, A. (2003), "A Deductive Variation on the No Miracles Argument", in *Synthese*, n. 201, pp. 81.

Hempel, C. G. (1952), *Fundamentals of Concept Formation in Empirical Science*, University of Chicago Press, Chicago, ILL.

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. (1986), *Induction: Processes of Inference, Learning and Discovery*, Cambridge, MA, MIT Press.

Howson, C. (2000), *Hume's problem: Induction and the justification of belief*, Oxford, The Clarendon Press.

Kitcher, P. (1993), *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*, New York, Oxford University Press.

Kitcher, P. (2001), "Real Realism: The Galilean Strategy", in *The Philosophical Review*, n. 110, 2, pp. 151-197.

Kosso, P. (1992), *Reading the Book of Nature: An Introduction to the Philosophy of Science*, Cambridge – New York, Cambridge University Press.

Laudan, L. (1981) "A Confutation of Convergent Realism", in *Philosophy of Science*, n. 48, 1, pp. 19-49.

Lyons, T.D. (2002), "The Pessimistic Meta-Modus Tollens", in Clarke, S. & T.D. Lyons (eds.), *Recent Themes in the Philosophy of Science. Scientific Realism and Commonsense*, Dordrecht, Kluwer.

Lyons, T.D. (2006), "Scientific Realism and the Stratagema de Divide et Impera", in The *British Journal for the Philosophy of Science*, n. 57, pp. 537– 560.

Magnus, P.D., and Callender, C. (2003), "Realist Ennui and the Base Rate Fallacy", in *Philosophy of Science* n. 71, pp. 320–338.

Morganti, M. (2024), "Alai on Novel Predictions and the Explanation of the Success of Scientific Theories", in Angelucci, A., Fano, V., Ferretti, G., Galli, G., Graziani, P., & Tarozzi, G. (eds.) (2024), pp.121-130.

Musgrave, A. (1988), "The Ultimate Argument for Scientific Realism", in Nola, R. (ed.) *Relativism and Realism in Science. Australasian Studies in History and Philosophy of Science*, vol 6, Dordrecht, Springer, pp. 229-252.

Perrin, J., (1913), *Les Atomes*, Paris, Alcan.

Psillos, S. (1999), *Scientific Realism. How Science Tracks Truth*, London, Routledge.

Stanford, PK. (2006), *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives: Science, History, and the Problem of Unconceived Alternatives*, New York, Oxford University Press.

Van Fraassen, B.C. (1980), *The Scientific Image*, Oxford, Clarendon Press.

Van Fraassen, B.C (2024) "Science does not describe reality. The limits and benefits of explanation", in *iaiNews*, 29th January. On-line: https://iai.tv/articles/science-does-not-describe-reality-auid-2724?_auid=2020&fbblid=IwAR0stKWef00M7UWyUZfY8hswgC2RhVrqH8lwxGege2bNf1PdrZt_2wRqkLM.

Vitruvius (1960), *The Ten Books on Architecture*, transl. by M.H. Morgan, Dover, New York.

White, R. (2003), The Epistemic Advantage of Prediction Over Accommodation, in *Mind*, n. 112 (448), pp. 653-683.

Wright, J. (2002), "Some Surprising Phenomena and Some Unsatisfactory Explanations of Them", in Clarke, S. & T.D. Lyons (eds.) *Recent Themes in the Philosophy of Science. Scientific Realism and Commonsense*, Dordrecht, Kluwer, pp.139–153.

Robots and Global Challenges: What we Need to Question for a More Sustainable Robotics

Ilaria Alfieri
IULM University, Milan
ilaria.alfieri@studenti.iulm.it

Antonio Fleres
IULM University, Milan
antonio.fleres@iulm.it

Maria Raffa
IULM University, Milan
maria.raffa@studenti.iulm.it

1. Introduction

This work lies at the intersection of sustainability and emerging technologies, a topic of increasing relevance. However, within the field of robotics, it is difficult to find literature that addresses sustainability by simultaneously considering the overall aspects of sustainability, i.e., the so-called three pillars, which refer to the economy, society and the environment as well (Cfr. Purvis *et al.*, 2019). Most existing research on sustainability in robotics, such as studies on green and soft robotics (Cfr. Hartmann *et al.*, 2021), focuses primarily on environmental concerns. These works examine issues like energy consumption, greenhouse gas emissions and pollution throughout the

Ilaria Alfieri, Antonio Fleres & Maria Raffa, "Robots and Global Challenges: What we Need to Question for a More Sustainable Robotics", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 53-74

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

lifecycle of robotic technologies, from production to disposal. In contrast, this paper takes a broader perspective, presenting three different issues that deserve attention not only in terms of the environmental dimension of sustainability.

The present paper is grounded on the evidence that in the debate about the global challenges facing our planet in these critical years, there are repeated calls for new technologies and artificial intelligence (AI) specifically. The European Commission's High Level Expert Group on Artificial Intelligence stated that AI is:

> a promising means to increase human flourishing, thereby enhancing individual and societal well-being and the common good, as well as bringing progress and innovation. In particular, AI systems can help to facilitate the achievement of the UN's Sustainable Development Goals, such as promoting gender balance and tackling climate change, rationalising our use of natural resources, enhancing our health, mobility and production processes (High-Level Expert Group on Artificial Intelligence, 2019: 4).

However, there is a tangible risk of exploiting AI and robotics for sustainable solutions to the environmental crisis without critically assessing their actual implications. This uncritical stance, which we may call "techno-enthusiasm" (Cfr. Coget, 2017), reflects an overly optimistic perspective on technology that overlooks its potential downsides. Techno-enthusiasm can lead to a failure to recognise and evaluate the disruptive effects of AI across various dimensions, social, ethical, political and environmental (Cfr. Floridi, 2022). Moreover, it can contribute to greenwashing, a specific form of ethical deception aimed at obscuring the environmental impacts of AI technologies and masking their true ecological consequences (Cfr. Heilinger *et al.*, 2023).

To explore these issues, this paper presents three key questions. Following this introduction, Section 2 examines the first question: is robot embodiment sustainable? It assesses whether current approaches to the physical design and embodiment of robots align with the needs of a sustainable human society. Section 3 deals with the second question: how can robotic embodiment be sustainable? It discusses Bayesian-based models as a potential framework for more sustainable robotic implementation. Section 4 investigates the third question: what role does social robotics play for sustainability? It reevaluates the role of social robotics in human-robot interactions, emphasizing their potential to promote both environmental and social sustainability. Finally, Section 5 presents the conclusions.

2. Is robotic embodiment sustainable?

The widespread adoption of robotics implies the extensive deployment of robotic embodiments. The concept of embodiment is broad and interdisciplinary, encompassing not only philosophy and psychology, but also fields such as psychology, communication, design and robotics (Cfr. Deng *et al.*, 2019). Considering the range of definitions of embodiment present in the literature, this paper adopts the one that frames it as the situated presence of a body – whether biological or artificial – within a specific environment (Cfr. Quick *et al.*, 2003). This definition is particularly relevant, as it provides the theoretical foundation used by scholars such as Feng, Dautenhahn, and Nourbakhsh to describe how embodiment enables robotic agents to participate in the human social sphere. A more detailed description of the definition of embodiment will be presented later in this section.

When it comes to the relationship between robotics and sustainability, one of the most important questions that we need to address is: is the physical embodiment of robots something that can be considered sustainable? This question leads to other relevant topics, such as: if so, is there a way to reduce the cost to human society and the environment? Hence, the question on the sustainability of robotic embodiment requires attention, both as a technological challenge and as an ethical responsibility. The relevance of this question arises for two primary rationales, summarised as: A) implication for environment and human society; B) its potential to reorient robotics toward responsible innovation.

A) Robotic embodiments have two distinct categories of implications – techno-ecological and societal – from a sustainability perspective. The first category concerns the environmental costs associated with the production, maintenance, and disposal of robotic embodiments. These processes contribute to systemic environmental challenges, including high energy consumption, depletion of rare earth materials, limited recyclability, and the generation of electronic waste (Cfr. Bugmann *et al.*, 2011). These costs are not just related to the technological compound of the robotic embodiment, but they involve even the external physical design. In particular, the use of plastics, paints, and synthetic materials in the construction of the robot – especially in social robots, where functionality is closely tied to physical appearance – can result in chemical pollution during manufacturing (Cfr. Fleres, 2025).

The second category is related to social implications, particularly those affecting social sustainability. The physical design of robotic embodiments can impact social dynamics, thus it is important to consider the effects on social sustainability. For example, Jennifer Robertson highlights the interplay between gender roles and robotic embodiment, showing how such design choices can perpetuate stereotypical representations and thereby contribute to gender-based inequalities and social disparities (Cfr. Robertson, 2017).

B) The second rationale is that this question is relevant for reorienting how design of robotics is conceived in relation to sustainability. It leads to the identification of opportunities for improvement and drives innovation that prioritises environmental protection and social responsibility. This requires rethinking what robotic embodiment should entail and whose values it ought to reflect. As robotics continues to evolve and integrate into various aspects of daily life, it is important to ensure that these technologies make a positive contribution to both the environment and society.

The importance of the concept of embodiment increased particularly in the 1990s, a period characterised by a profound re-evaluation of the contribution of the physical body to cognitive processes. This shift fundamentally changed fields such as cognitive science, artificial intelligence and robotics as well, by emphasising the crucial role of the body in shaping cognition (Cfr. Damiano & Dumouchel, 2020). As previously noted, many researchers have proposed different definitions of embodiment in robotics, which are worth mentioning in more detail. Pfeifer and Scheier describe embodiment in the following terms:

> A term used to refer to the fact that intelligence cannot merely exist in the form of an abstract algorithm, but requires a physical instantiation, a body. In artificial systems, the term refers to the fact that a particular agent is realised as a physical robot or as a simulated agent (Pfeifer & Scheier, 2001: 649).

This definition highlights the importance of the presence of a physical body in order to have an intelligent artificial system that can act in a physical space. Thus, the term "embodiment" here refers exclusively to the physical presence of the robot itself. A different focus on embodiment is given by Quick and colleagues:

> A system X is embodied in an environment E if perturbatory channels exist between the two. That is, X is embodied in E if for every time t at which both X and E exist, some subset of E's possible states have the capacity to perturb X's state, and some subset of X's possible states have the capacity to perturb E's state (Quick et al., 2003: 653).

As stressed at the beginning of the section, this definition brings the environment into focus, showing how it contributes to embodiment by enabling a structural coupling between the agent and its environment.

Considering this, it can be summarised that robotic embodiment is a key feature that enables artificial agents to act autonomously in the physical world. It supports crucial abilities such as environmental integration, manipulation, navigation and sensory feedback – much like human perception aids decision-making (Cfr. Mergner *et al.*, 2019). In social robotics, embodiment also shapes human-robot interaction by influencing social presence and engagement (Cfr. Dumouchel & Damiano, 2017). Researchers like Dautenhahn and colleagues (2002) have emphasized the importance of *social embeddedness*, where a robot's physical form allows it to become part of a social system through structural coupling. Studies have shown that physical embodiment grants to artificial agents a more efficient social presence (Cfr. Heerink *et al.*, 2010; Shinozawa *et al.*, 2002). Despite sustainability concerns, physical embodiment remains essential, not only for interaction and intelligence (Cfr. Brooks, 1991), but also for operating in the real world's complexity.

While green and soft robotics have explored aspects of sustainability (Cfr. Shintake, 2022), the embodiment of robots remains largely unexamined in this regard. The global challenge of sustainability requires us to not only focus on the environmental aspects of sustainability, but to tackle all three pillars of sustainability. The embodiment of robots provides an excellent starting point to address the problem of sustainability.

Although the physical embodiment of social robots represents a critical nexus linking humans, society and the environment, it is important to stress that a physically embodied robotic agent constitutes only one point within a broader and more diverse spectrum of embodiment possibilities. Indeed, for instance, the term "virtual embodiment" (Cfr. Deng *et al.*, 2019) is often applied to artificial agents displayed on screens or to body parts represented digitally, such as virtual faces. Furthermore, with the advent of augmented reality, mixed-reality embodiments have emerged, where some components are physical while others exist virtually and are visualized through specialized headsets (Cfr. Dragone *et al*., 2009).

Having clarified the centrality of embodiment in robotics, it can now be explorex more in detail how it can result in concrete practices of (un)sustainability in real-world contexts, with both techno-ecological and societal implications.

Firstly, concerning the techno-ecological implication, the impact of the embodiment of robots on the environment is significant. The physical embodiment of a robot influences its interaction with the environment. It includes aspects such as the manipulation of physical objects, energy consumption, material requirements and maintenance needs. The robot's operational activities, including its interaction with and modification of objects, reflect a dynamic engagement with its environment. Furthermore, the constant need for energy and maintenance underscores the robot's ongoing impact on the material world.

Secondly, considering the societal implication, the emotional and social dimensions of robotic embodiment are closely intertwined and have significant implications for how social robots interact with individuals and communities. The emotional aspect of robotic embodiment is evident in the interactions between social robots and their users. These robots are specifically designed to foster emotional connections and elicit anthropomorphic responses that can significantly influence personal behaviour and dynamics in private environments (Cfr. Damiano & Dumouchel, 2018; Fink, 2012). The notion of artificial empathy, as explored by Dumouchel & Damiano (2017), demonstrates how robots can engage in affective synchronisation with humans, enabling nuanced emotional exchanges and increasing the depth of human-robot interactions. Furthermore, the embodiment of robots has profound implications for society. The design and physical characteristics of social robots play a critical role in their integration into broader societal systems such as labour markets and urban environments. The way in which robots are embodied has a direct impact on their functionality and social role, affecting both their operational efficiency and their acceptance in different social contexts. The alignment between the embodied form of a robot and its intended societal function is critical to its effectiveness and the degree to which it is accepted in different societal contexts.

Thirdly, considering both techno-ecological and societal implications, the economic perspective on the embodiment of robots must also be considered. The development and maintenance of robots involve significant costs, which raises concerns about economic sustainability. The high costs associated with robotic technology may exacerbate inequalities and create a divide between individuals who can afford the benefits of robotics and those who cannot. This potential economic gap highlights the need to assess the broader financial impact of robotics on social justice. In other words, the interconnected dimensions of environmental impact, emotional engagement, economic considerations, and societal impact emphasise the importance of robotic embodiment in shaping interactions and effects across multiple domains. This comprehensive understanding highlights the intricate relationships and feedback mechanisms that arise from the physical presence of robots, allowing sustainability to be viewed from a new, broader and transdisciplinary perspective.

Hence, addressing the concept of *physical embodiment* is particularly crucial when considering sustainability. The term should not be conflated with *robotic body*, as the two are neither synonymous nor reducible to the same construct. As discussed in this section, embodiment is intrinsically linked to the environment in which the robot operates. In the context of social robotics, *physical embodiment* exists within a social environment, wherein human-robot interactions are fundamentally mediated by the robot's embodied presence. The design and characteristics of this embodiment significantly influence the nature and quality of social interactions with human users. For this reason, it is imperative to examine the sustainability of *physical embodiment* rather than solely focusing on the material sustainability of the robot's body. Reducing embodiment to a single dimension – such as material or energy consumption – would fail to capture its broader implications. Notably, social robots entail sustainability costs beyond environmental and economic factors, extending into social dimensions that affect both present and future human communities. Consequently, achieving a holistic model of sustainability – one that fully integrates economic, social, and environmental considerations – necessitates a shift away from a narrow focus on the robot's body and instead demands a comprehensive analysis of embodiment in its full complexity.

Therefore, addressing the sustainability of robotic embodiment is a necessary step in reorienting the field towards a more sustainable and equitable future. On this basis, in the next section we will investigate the relationship among sustainability and robotics by addressing the issue of the implementation of robots.

3. How to make the implementation of robots sustainable?

In the previous section, we examined the physical embodiment of robots and assessed whether current design approaches align with sustainability principles. Now, we turn to a complementary aspect: the computational models underlying robotic implementation. Specifically, we explore how certain frameworks can contribute to sustainability in distinct ways.

As discussed, robot embodiment presents several environmental challenges, particularly regarding material disposal and energy consumption. From a technical perspective, the first step toward addressing these issues is to enhance the energy efficiency of robotic implementation, optimizing the algorithmic processes that drive these systems to reduce power consumption. However, as we have seen above, sustainability extends beyond ecological concerns. It is equally important to consider social sustainability, particularly in the way AI systems interact with humans. Indeed, a key aspect of socially sustainable AI is explainability – the ability of an AI system to make its decision-making processes transparent to both developers and users. Enhancing explainability fosters trust and accountability, ensuring that robotic technologies are not only efficient but also ethically and socially responsible (Cfr. Mazzi & Floridi, 2023; Heilinger *et al.*, 2023).

One computational approach that intersects these concerns is Predictive Coding (PC). The PC, or predictive processing, is an influential theory in computational and cognitive neuroscience, proposing that the core function of the brain is to minimise prediction errors, i.e., signal mismatches between predicted input and the input actually received from the environment. This minimisation can be achieved in a number of ways:

> Through immediate inference about the hidden states of the world, which can explain perception, through updating a global world-model to make better AI predictions, which could explain learning, and finally through action to sample sensory data from the world that conforms to the predictions (Millidge *et al.*, 2022: 3).

In other words, the PC can be seen as a unified account of perception, action and cognition, in which the brain is seen as a predictive machine that tries to predict its next states on the basis of the information gained from the previous interaction with the environment (Cfr. Friston *et al.*, 2011). This means that the brain always tries to minimise the probability of prediction errors and aims to avoid high surprise states. PC can be described as an approximate Bayesian inference process based on Gaussian inference (Cfr. Millidge *et al.*, 2022). Andy Clark hypotheses that PC may also be useful for understanding imaginative processes, since, in Clark's words, perceivers are also imaginators (Cfr. Clark, 2016). Furthermore, free energy minimisation falls under the broader umbrella of the Free Energy Principle (FEP), theorized by Karl Friston, that can be understood as a general methodology for optimising resources within an agent or system. FEP is based on the premise that systems are separated from their environment but interact with it through a statistical boundary known as the Markov blanket. FEP states that random dynamical systems that are coupled to but separate from each other will appear to track or infer each other's behaviour, and that agents that exist will do so because they can persist and maintain their equilibrium through free energy minimisation. Free energy can be written in different ways:

> First, it can be expressed as expected energy minus the entropy of the variational density, which licenses the name *free energy*. In this decomposition, minimising variational free energy corresponds to the maximum entropy principle, under the constraint that the expected free energy is minimised [...]. Second, variational free energy can be decomposed into the (negative) log likelihood of particular states (i.e. negative *accuracy*) and the KL divergence between posterior and prior densities (i.e. *complexity*). Finally, it can be written as the self information associated with particular states (i.e. *surprisal*) plus the KL divergence between the variational and posterior density, which is zero (Friston *et al.*, 2023: 17).

Free energy minimisation is achieved through active inference (AIF), which – again – is a process that uses information from the history of previous interactions with the environment to modify current states and suppress errors in predicting future states (Cfr. Friston *et al.*, 2011; Kirchhoff, 2018). In a very intuitive claim, Friston describes AIF as "feeling our way in the dark, anticipating what we might touch next, and then trying to confirm those expectations" (Friston, 2010: 129).

AIF is an attractive framework for implementing robotic applications where the robot or task dynamics are uncertain. For estimation, adaptive control, fault-tolerant control, prospective planning, and complex cognitive abilities (human-robot cooperation, self/other discrimination) (Cfr. Lanillos *et al.*, 2021). In addition, the implementation of the PC with deep neural networks has gained popularity in the computer vision community for modelling multisensory perception and for video prediction.

While Friston was developing the foundations of FEP and AIF, Jun Tani and colleagues were investigating models similar to AIF in real robots, by showing that a robot could successfully adapt its movement pattern to the appropriate movement primitive in real time as the environment changed (Cfr. Tani, 2003). However, these models were still limited because they were based on a deterministic dynamics perspective rather than the Bayesian perspective used in the formal formulation of AIF (Cfr. Friston *et al.*, 2011). A robotic trial of AIF with Friston's exact formalism for reaching tasks was then performed with a 7-DOF simulated robotic arm with the generative models and parameters known in advance (Cfr. Pio-Lopez *et al.*, 2016).

On this path, Pablo Lanillos and Gordon Cheng (2018) implemented a computational model to enable a robot to infer its own body configuration. In this model, PC is used for a computational perceptual model that allows any multisensory humanoid robot to learn, infer and update its own body configuration. This model allows generic multisensory integration by integrating different sources of information (tactile, visual and proprioceptive): the robot estimates and adjusts its body configuration using only sensory information. In this sense, AIF is well suited to model decision making. In fact, AIF-based robots model the intentions of others to predict their actions, such as movements, thus enabling intentional understanding. This allows robots to operate safely in social environments by constantly resolving uncertainty about others' intentions and implicit goals. This embodiment is particularly crucial for social assistive robotics, such as personal assistants, robotic nurses and companions, e.g. for assisting the disabled and elderly (Cfr. Da Costa *et al.*, 2022).

Taking all this into account, it is clear that while PC and AIF were originally developed to explain biological cognition, their application to robotics has potential sustainability benefits in both social and environmental domains.

From a social sustainability perspective – understood as the ability of robotic systems to integrate into the human environment – these frameworks contribute to safety, adaptability and transparency. AIF-based agents continuously resolve uncertainty by selecting informative actions that minimize risk, a crucial factor in high-stakes, unpredictable scenarios such as human-robot interaction. By reducing ambiguity and optimizing decision-making, these models enable robots to anticipate and respond to dynamic environments in a more reliable manner. Moreover, when faced with uncertainty, AIF-driven robots can autonomously seek guidance from users, for instance, through shared control mechanisms (Cfr. Da Costa *et al.*, 2022). This enhances human-robot collaboration and increases operational transparency, strengthening trust in robotic systems.

AIF also promotes explainability, an essential aspect of socially sustainable AI. Unlike complex black-box models (e.g., deep learning networks based on feedforward architectures), AIF is grounded in Bayesian networks, which follow explicit causal reasoning. This characteristic makes AI decisions more transparent and understandable, reinforcing accountability and ethical AI practices (Cfr. Albarracin *et al.*, 2023). In this sense, AIF aligns with the principles of sustainable AI ethics, as transparency and traceability are key factors in building ethical and socially sustainable systems (Cfr. Mazzi & Floridi, 2023; Van Wynsberghe, 2021).

Additionally, AIF enhances adaptability in changing environments. By dynamically adjusting actions based on evolving knowledge, it ensures that robotic decision-making is robust and context-aware, balancing short-term and long-term objectives. This adaptability is particularly important in real-world applications where conditions and requirements evolve rapidly.

Beyond social sustainability, AIF also contributes to ecological sustainability by improving energy efficiency. By optimizing action selection, reducing unnecessary movements, and prioritizing information-seeking behaviors that lower computational costs, AIF helps mitigate excessive energy consumption, making robotic operations more resource-efficient.

As mentioned above, AIF models are closely tied to the FEP, which provides a very general framework for resource optimisation. Indeed, the FEP is a comprehensive theory that aims to explain how biological systems maintain their internal states by minimising the discrepancy between predicted and actual sensory inputs. This principle has been applied to a wide range of systems, from neural networks to organisational structures, demonstrating its versatility. However, it is important to consider Bayesian-based models' weaknesses. These models are based on basic decision theory, which assumes an optimal decision maker. This theoretical decision maker is assumed to have the ability to calculate and choose the move that maximises the utility function at each stage of problem solving. Moreover, these theories of expected utility maximisation have been criticised for being computationally intractable, especially when dealing with systems involving a large number of random variables. As Johan Kwisthout and Iris van Rooij (2020) point out, the computational complexity of such models grows exponentially with the number of states. This results in a significant increase in computational effort and energy consumption, making it difficult to apply these models efficiently in real-world robotic systems. The high energy demands of FEP-based models pose concerns about their environmental impact when implemented at scale.

Thus, while AIF provides a strong theoretical foundation for sustainable resource allocation and energy-efficient action selection, its computational cost remains a major limitation. Balancing efficiency and sustainability remains a critical challenge, underscoring the need for further research and development to enhance the energy efficiency of these models.

In summary, AIF offers potential contributions to both social and ecological sustainability for robotic implementation. It enhances human-robot interaction, transparency, and explainability, promoting trust and ethical AI practices. At the same time, it provides a framework for energy-efficient robotic operations, though its computational costs must be carefully managed.

So far, this paper has examined robotic embodiment and implementation, addressing both environmental and social sustainability concerns. The next section continues to explore social sustainability, shifting the focus to human-robot interactions and proposing the third and final question: *what role does social robotics play in sustainability*?

4. What role does social robotics play in sustainability?

As robotics technology continues to evolve and spread in our society, the concept of social robots – robots designed to interact with humans in a socially meaningful way – has gained increasing attention. According to Korn, "social robots are robots which cannot only do services for us but also communicate – thus, they could come very close, into our homes, into our private lives" (Korn, 2019: V) becoming real "social partners" (Cfr. Dumouchel & Damiano, 2017) to interact with. One emerging field within this domain is "social robotics for sustainability" (Cfr. Alfieri, Fleres, Damiano, 2022), which can be defined as the application of social robotics technologies, exploiting multimodal communication modes based on social cues (e.g. emotions, body language), to promote sustainable behaviour among users. Within this new research direction one interesting option of development is related to one particular approach: Persuasive Social Robotics. A persuasive social robot is an embodied agent (robot) that can interact socially with humans and significantly influence or change their behaviour, attitudes, or cognitive processes (Cfr. Siegel, 2009). They implement this change by using persuasive strategies in their interaction with humans. The purpose of persuasive social robots is thus to harness their social power to direct humans towards goals that are relevant to those who design and produce them. Therefore, the design choices of these technologies are extremely important, because through persuasive power users can be directed towards certain behaviours rather than others, and these can have positive or negative consequences. Examples of persuasive social robots might include those robots programmed with the aim of encouraging exercise and preventive gymnastics in the elderly (Cfr. Tanioka *et al.*, 2019); robots in healthcare that persuade patients to adhere to a specific therapeutic programme or that deal with health care in general (Cfr. Looije, 2010); robots that motivate users to lose weight (Cfr. Kidd & Breazeal, 2007); tutor robots that persuade children to learn to do their homework (Cfr. Ham *et al.*, 2011), assistive social robots that attempt to negotiate an activity schedule with their user (Cfr. Ficocelli *et al.*, 2016).

As mentioned before, persuasion can be achieved by implementing different strategies in the robot. Such as providing social feedback whether positive or negative – that can persuade the user to have one behaviour instead of another (Cfr. Midden & Ham, 2009), or gratification. Further is the strategy identified by Augello et al., in which the use of narrative arguments, such as storytelling, together with the use of the user's emotional responses, persuades people to vaccinate against Covid 19 (Cfr. Augello *et al.*, 2021). In addition, robots should use social influence strategies. Studies have also shown that the persuasive effect increases if the robot customises its responses according to the specific user it interacts with and their needs, using the engaging and empathic aspects of persuasion (Cfr. Saunderson & Nejat, 2020). These strategies can be designed to encourage certain behaviours, persuade towards a certain line of thinking, or convince individuals to act in a certain way. For instance, there are several studies that suggest that persuasive social robots can use their persuasive influence to steer users towards more sustainable attitudes towards the environment. Indeed, the persuasive nature of a robot can have positive effects on encouraging pro-environmental behaviour (Beheshtian *et al.*, 2020). Persuasive social robots can help reduce energy consumption (Cfr. Ham & Midden, 2014), help improve children's waste separation practices (Cfr. Castellano *et al.*, 2021), implement pro-environmental/sustainable behaviour (Cfr. Tussyadiah & Miller, 2019) and encourage sustainable behaviour in shared living spaces (Cfr. Beheshtian *et al.*, 2020). The creation of these scenarios represents a purely experimental stage of development. However, they can provide insights into the potential applications of persuasive social robots in encouraging pro-environmental behaviours. Furthermore, they can inform the direction of future research and development efforts, guiding the design of persuasive technologies towards the promotion of specific pro-environmental actions.

Nevertheless, while the purpose of these robots is praiseworthy, a more critical analysis of the role of these robots in sustainability has shown us a series of limitations, especially of an ethical nature, which cannot be ignored when we discuss the interaction between humans and robots. Indeed, the use of persuasion in social robots poses several ethical issues such as manipulation, interference with autonomy of the users, acceptance of the robot, psychological reactance, asymmetrical persuasion, user awareness of persuasion etc. These concerns are particularly relevant in the context of sustainability, where the goal should be to encourage voluntary behaviour change rather than impose it. Persuasion involves exerting influence on someone but without the use of coercion or deception (Cfr. Fogg, 2003).

Despite this, persuasion and manipulation are frequently confused or linked. From Cambridge Dictionary, manipulation is "the action of influencing or controlling someone or something to your advantage, often without anyone knowing it". Indeed, one of its main characteristics is that the purpose of manipulation must be concealed. Additionally, according to Breton's definition, manipulation is "a violent and restrictive action that deprives those who are subjected to it of their freedom" (Nettel & Roque, 2012: 59). That is why we should be very careful when using these robots, because their persuasive influence can persuade users to behave in a way they might not independently choose. There can be a fine line between persuasion and manipulation. Any alteration of habits and behaviour must be undertaken on a personal level, initiated by the individual rather than by external influences such as robots that dictate right and wrong. Such influence could potentially interfere with the autonomy of the individual. Very generically, "autonomy is self-determination: the ability to do what one does independently, without being forced to do so by some outside power" (Boden, 2008: 305). When robots are designed to persuade, there is a risk of an outside power that might undermine this autonomy by subtly influencing decisions and actions, thereby limiting the user's freedom to choose. It is important to clarify that this reflection does not concern science-fictional scenarios in which robots might coercively impose behaviours on users, and force them to do things, thus limiting their autonomy. This is evidently unrealistic in current technological contexts. Rather, the concern lies in how social robots, through their design and persuasive strategies, can subtly shape user behaviour, sometimes in ways that may escape the user's full awareness. For instance, while a robot may encourage healthy eating and the practice of daily exercise, and such suggestions are obviously non-coercive, their repeated presentation, emotional framing, or personalization may nevertheless influence users toward predetermined behavioural patterns. In this sense, ensuring that the user retains a clear sense of agency and the ability to critically assess or decline the robot's suggestions remains an ethically significant design objective.

Furthermore, there is the issue of human acceptance of these technologies. Users may react negatively to a robot telling them what to do, which could lead to the phenomenon of "psychological reactance" (Cfr. Ghazali *et al.*, 2018) such as negative feelings and emotions that make it difficult for them to follow the persuader's advice. This could potentially compromise the quality of the interaction. Persuasive social robots, when perceived as controlling or intrusive, might trigger such reactance, resulting in the opposite of the desired outcome. This could manifest as users rejecting sustainable behaviours due to a sense of pressure.

A further essential ethical concern is the transparency of the robot's actions (Cfr. Wortham, 2020) and the necessity for the user to be constantly aware of the persuasive techniques employed by the robot. In order for persuasion to be ethically acceptable, users must be aware that they are being persuaded and must have the option to decline. However, social robots may employ subtle or implicit techniques that users are not fully aware of, resulting in behaviour changes without informed consent. This lack of awareness undermines the ethical principle of transparency.

Another significant challenge for maintaining ethically human-robot interactions when using persuasive social robots is the fact that persuasive strategies, that usually are acceptable in a human-human context, might be considered ethically problematic in human-robot context. This is primarily due to the inherent asymmetry between the persuader and the persuaded (Cfr. Nickel & Spahn, 2012). In human-robot interaction, the robot in that moment possesses the capacity to influence the user, but the user, conversely, cannot exert the same level of influence over the robot, at least during the course of that specific interaction. Unlike human counterparts, the robot's behaviour and responses are pre-programmed, meaning the user cannot prompt changes in the robot's conduct during the interaction. This imbalance highlights a significant ethical issue, as it places the robot in a position of persuasive power without reciprocal agency from the user.

The brief analysis of the limitations of persuasive social robotics has led us to reflect on whether this is the best approach and role to use social robotics for sustainability. Limiting the contribution of social robotics to mere persuasion represents an overly narrow and potentially misleading approach. The challenges of sustainability are inherently complex and multidimensional and cannot be effectively addressed solely by attempting to change individual habits. An approach focused exclusively on persuasion risks ethical and behaviourist drifts. Rather, it is necessary to critically and comprehensively rethink another role that social robotics can and should play within a sustainability framework. A rethinking of the role of social robotics for sustainability requires moving beyond the assumption that these robots are *a priori* solutions to sustainability challenges. It also means to recognize that guiding human behaviour through persuasive social robots is not the only means to promote sustainability. Instead, firstly it is essential to gain a deeper understanding of how social robots can be used to address these issues and to identify ways in which they can be deployed in a sustainable manner, without generating new ethical or social concerns. Secondly, social robotics should adopt a complex approach with the aim of achieving sustainability in all three dimensions: environmental, social, and economic. In light of the aforementioned considerations, we propose to shift the focus from influence to partnership, imagining a new generation of social robots no longer designed to direct user behaviour, but to co-construct sustainable relational dynamics, both socially and environmentally. By "partnership" we mean a relationship of collaboration and coordination between humans and robots, in which the robot is not merely a tool, but acts as a dynamic partner, capable of supporting the individual in an ethical and social manner.

In this perspective, the social robot is not influencing the user's behaviour through persuasive strategies, but is configured as a situated partner, cooperating within everyday contexts to promote conditions favourable to the emergence of sustainable lifestyles. This type of robot can contribute, for example, to forms of sustainable living – both at the level of individual and community well-being (e.g., care, assistance, quality of life, social cohesion) and of collective ecological responsibility (e.g., shared environmental tasks, community engagement, local sustainability initiatives). By integrating these ecological and social aspects, the role of social robots should be oriented towards enhancing people's quality of life. Linked to the quality of life is the concept of well-being. Indeed, research in the field of social sustainability indicates that the concept of well-being is of central importance, being inextricably linked to the quality of life and the relationship between humans and their natural environment (Cfr. Bandarage, 2013; Helne & Hirvilammi, 2015). Therefore, a social robot designed for sustainability should not only promote ecologically responsible behaviour but also contribute to the enhancement of people's quality of life by encouraging a harmonious relationship between individuals and the ecosystem. In conclusion, in this last section we presented another challenge of social robotics for sustainability that we must be prepared to embrace and advance in the near future. That is, not merely about ensuring the sustainability of robot's embodiment and implementation; it is also about considering the role they play and the interactions they engage in with humans. By embracing this challenge, we can advance the field of social robotics in a way that makes a meaningful contribution to a more sustainable and equitable future.

5. Conclusion

Robotics represents one of the most significant global challenges emerging from technological progress. While it has the potential to serve as a valuable tool for enhancing sustainability on Earth, it also presents risks that could have detrimental consequences. This paper has tried to provide a comprehensive analysis of this dual challenge, focusing on three dimensions of both environmental and social sustainability: embodiment, implementation and human-robot interaction. This has been done by interrogating embodiment, implementation and interaction.

Specifically, first it was asked: is robotic embodiment sustainable? The question revealed that while physical embodiment is essential for robotic functionality and social interaction, it raises significant sustainability concerns across environmental, economic and social domains. Addressing these requires a shift toward an evaluation of embodiment beyond mere material impact.

Second, the paper examined how to make the implementation of robots sustainable. In that section, AIF was explored as a promising framework for that purpose. Indeed, AIF offers benefits for social and ecological sustainability through adaptability, explainability and energy efficiency – though its high computational demands remain a challenge that must be addressed.

Third, the paper explored the role of social robotics for sustainability. It was argued that the prevalent focus on persuasive social robots risks ethical pitfalls, such as manipulation and diminished autonomy. Instead, it was proposed to rethink social robots as partners in co-creating sustainable relational dynamics that foster well-being and ecological responsibility.

Although this work does not claim to offer definitive solutions to these complex challenges, it aims to outline a path forward. By highlighting the need to include sustainable robotics into the broader discourse on global challenges, this paper emphasises the importance of adopting a holistic and interdisciplinary perspective. Sustainable robotics must be considered within the interconnected framework of environmental, social and economic sustainability to ensure that technological advances contribute positively to the future of our planet.

References

Albarracin, M., Hipolito, I., Tremblay, S. E., Fox, J. G., René, G., Friston, K., & Ramstead, J. D. M. (2023), “Designing explainable artificial intelligence with active inference: A framework for transparent introspection and decision-making”, in *ArXiv*, ArXiv 2306.04025.

Alfieri, I., Fleres, A., & Damiano, L. (2022), Workshop "Eco-socio-botics 2022 - Social Robotics for Sustainability" at 14th International Conference, ICSR 2022, Florence, Italy, December 13-16, 2022.
Augello A., Città, G., Gentile M., & Lieto, A. (2021), "A Storytelling Robot Managing Persuasive and Ethical Stances via ACT-R: An Exploratory Study", in *International Journal of Social Robotics*, 1-20.
Bandarage, A. (2013), *Sustainability and well-being: The middle path to environment, society and the economy*, Palgrave Macmillan, London.
Beheshtian, N., Moradi, S., & Ahtinen, A. (2020), "GreenLife: A Persuasive Social Robot to Enhance the Sustainable Behavior in shared Living Spaces", in *NordiCHI '20: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society,* n. 42, pp. 1-12.
Boden, M. A. (2008), "Autonomy: What is it?" in *Biosystems*, 91, n. 2, pp. 305-308.
Brooks, R. A. (1991), "Intelligence without representation" in *Artificial Intelligence*, 47, n. 1-3, pp. 139-159.
Bugmann, G., Siegel M., & Burcin, R. (2011), "A Role for Robotics in Sustainable Development?", in *IEEE AFRICON Conference*, pp. 1-4.
Castellano, G., De Carolis, B., & D'Errico, F. (2021), "PeppeRecycle: Improving Children's Attitude Toward Recycling by Playing with a Social Robot", in *International Journal of Social Robotics*, n. 13, pp. 97-111.
Clark, A. (2016), *Surfing uncertainty: Prediction, action and the embodied mind*, Oxford, Oxford University Press.
Coget, J. (2017), "Technophobe vs. Techno enthusiast: Does the Internet Help or Hinder the Balance Between Work and Home Life?", in *Academy of Management Perspectives*, 25(1), pp. 95-96.
Da Costa, L., Lanillos, P., Sajid, N., Friston, K. J., & Khan, S. (2022), "How Active Inference Could Help Revolutionise Robotics", in *Entropy*, n. 24, p. 361.
Damiano, L., & Dumouchel, P. (2020), "Emotions in relation. Epistemological and ethical scaffolding for mixed human-robot social ecologies", in *Humana Mente*, 13, n. 37, pp.181-206.
Dautenhahn, K., Ogden, B., & Quick, T. (2002), "From embodied to socially embedded agents - Implications for interaction-aware robots" in *Cognitive Systems Research*, 3, n. 3, pp. 397-428.
Deng, E., Mutlu, B., & Mataric, M. J. (2019), "Embodiment in Socially Interactive Robots", in *Foundations and Trends in Robotics*, 7, n. 4, pp. 251-356.

Dragone, Mauro, Thomas Holz, G. M. P. O'Hare, & Michael J. O'Grady, "Mixed Reality Agent (MiRA) Chameleons", 2009, 13-33.

Dumouchel, P., & Damiano, L. (2017), *Living with Robots*, Harvard University Press (*Vivere con i Robot. Saggio sull'Empatia Artificiale*, tr. L. Damiano, Raffaello Cortina Editore, 2019).

Ficocelli, M., Terao, J., & Nejat, G. (2016), "Promoting interactions between humans and robots using robotic emotional behavior", in *IEEE Transactions on Cybernetics*, 46, n. 12, pp. 2911-2923.

Fleres, A. (2025), *Robotica (sociale) e sostenibilità. Impalcature teoriche della robosfera,* PhD Thesis, IULM University.

Floridi, L. (2022), *Etica dell'Intelligenza Artificiale: Sviluppi, Opportunità, Sfide*, Milano, Raffaello Cortina Editore.

Fogg, B. J. (2003), *Persuasive Technology. Using Computers to Change What We Think and Do*, San Francisco, CA, Morgan Kaufmann Publishers.

Fong, T., Nourbakhsh, I., Dautenhahn, K. (2003), "A Survey of Socially Interactive Robots: Concepts, Design, and Applications", in *Robotics and Autonomous Systems*, 42, n. 3-4, pp. 143-166.

Friston, K. J. (2010), "The Free Energy Principle: A Unified Brain Theory?", in *Nature Reviews Neuroscience*, 11, n. 2, pp. 127-138.

Friston, K. J., Mattout, J., & Kilner, J. (2011), "Action Understanding and Active Inference", in *Biological cybernetics*, 104, n. 1, pp. 137-160.

Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G.A., & Parr, T. (2023), "The Energy Principle Made Simpler but not Too Simple", in *Physics Reports*, 1024, pp. 1-29.

Ghazali, A. S., Ham, J., Barakova E. I., & Markopoulos, M. (2018), "Poker Face Influence: Persuasive Robot with Minimal Social Cues Triggers Less Psychological Reactance" in *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication,* pp. 940-946.

Ham, R., Bokhorst, R., Cuijpers, H., Van Der Pol, D., & Cabibihan, J. J. (2011), "Making robots persuasive: The influence of combining persuasive strategies (gazing and gestures) by a storytelling robot on its persuasive power", in *Proc. Int. Conf. Soc. Robot.*, 2011, pp. 71-83.

Ham, J., & Midden, C. (2014), "A Persuasive Robot to Stimulate Energy Conservation: The Influence of Positive and Negative Social Feedback and Task Similarity on Energy-Consumption Behavior", in *International Journal of Social Robotics*, 6, pp. 163-171.

Hartmann, F., Baumgartner, M., & Kaltenbrunner, M. (2021), "Becoming Sustainable, The New Frontier in Soft Robotics", in *Advanced Materials*, 33, n. 19, 2004413.

Heerink, M., Kröse, B., Evers, V., Wielinga, B., Heerink, M., Kröse, B., Evers, V., Wielinga, B., Evers, V., & Wielinga, B. (2010), "Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model", in *International Journal of Social Robotics*, 2, pp. 361-375.

Heilinger, J., Kempt, H., & Nagel, S. (2023), "Beware of sustainable AI! Uses and abuses of a worthy goal", in *AI Ethics*, pp. 1-12.

Helne, T., & Hirvilammi, T. (2015), "Wellbeing and Sustainability: A Relational Approach", in *Sustainable Development*, 23, n. 3, pp. 167-175.

High-Level Expert Group on Artificial Intelligence (2019), *Ethics guidelines for trustworthy AI*, B-1049 Brussels.

Kidd, C. D., & Breazeal, C. (2007), "A Robotic Weight Loss Coach", in *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, Menlo Park, California.

Kirchhoff, M. D. (2018), "Predictive Processing, Perceiving and Imagining: Is to Perceive to Imagine, or Something Close to it?", in *Philosophical Studies*, 175, n. 3, pp. 751-767.

Korn, O. (2019) *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction*, Offenburg, Springer.

Kwisthout, J., & van Rooij, I. (2020), "Computational resource demands of predictive Bayesian brain", in *Comput. Brain Behav.* 3, pp. 174-188.

Lanillos, P., & Cheng, G. (2018), "Adaptive robot body learning and estimation through predictive coding2, in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 4083-4090.

Lanillos, P., Meo, C., & Pezzato, C. (2021), "Active inference in robotics and artificial agents: survey and challenges", in *ArXiv*, arXiv:2112.01871.

Looije, M., Neerincx A., & Cnossen F. (2010), "Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors", in *International Journal of Human-Computer Studies*, 68, n. 6, pp. 386-397.

Mergner, T., Funk, M., & Lippi, V. (2019), "Embodiment and Humanoid Robotics", in *Philosophisches Handbuch Künstliche Intelligenz*, 1, pp. 1-27.

Mazzi F., & Floridi L. eds. (2023), *The Ethics of Artificial Intelligence for the Sustainable Development Goals*, Cham, Switzerland, Springer.

Midden, C., & Ham, J. (2009), "Using negative and positive social feedback from a robotic agent to save energy", in *Proceedings of the 4th International Conference on Persuasive Technology. Association for Computing Machinery*, New York, USA.

Millidge, B., Seth, A., & Buckley, C. (2022), "Predictive Coding: A Theoretical and Experimental Review", in *ArXiV*, arXiv: 2107.2979.

Nettel, A. L., & Roque, G. (2012), "Persuasive Argumentation Versus Manipulation", in *Argumentation*, 26, pp. 55-69.

Nickel, P., & Spahn, A. (2012), "Trust, discourse ethics, and persuasive technology" in *Proceedings of the 7th International Conference on Persuasive Technology, Persuasive 2012, Linköping University Electronic Press,* pp. 37-40.

Quick, T., Dautenhahn, K., Nehaniv, C. L., & Graham, R. (2019), "The Essence of Embodiment: A Framework for Understanding and Exploiting Structural Coupling between System and Environment", in *AIP Conference Proceedings (AIP, 2003)*, pp. 649-660.

Pfeifer, R., & Scheier, C. (2001), *Understanding Intelligence*, Cambridge, Massachusetts, MIT Press.

Pio-Lopez, L., Nizard, A., Friston, K. J., & Pezzulo, G. (2016), "Active inference and robot control: a case study", in *Journal of The Royal Society Interface*, 13, n. 122.

Purvis, B., Mao, Y., & Robinson, D. (2019), "Three pillars of sustainability: in search of conceptual origins", in *Sustainability Science*, 14, pp. 681-695.

Saunderson, S., & Nejat, G. (2020), "Investigating Strategies for Robot Persuasion in Social Human-Robot Interaction", in *IEEE Transactions on Cybernetics*, 52, n. 1, pp. 641-653.

Shinozawa, K., Reeves, B., Wise, K., So-Hye, L., Maldonado, H., & Naya, F. (2003), "Robots as New Media: A Cross-Cultural Examination of Social and Cognitive Responses to Robotic and On-Screen Agents", in *Proceedings of annual conference of international communication association*, pp. 998-1002.

Shintake, J. (2022), "Green Robotics: Toward Realization of Environmentally Friendly Soft Robots", in *Journal of Robotics and Mechatronics*, pp. 270-272.

Siegel, M., Breazeal, C., & Norton, M. (2009), "Persuasive Robotics: The influence of robot gender on human behavior", in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2563-2568.

Tani, J. (2003), "Learning to generate articulated behavior through the bottom- up and the top-down interaction processes", in *Neural networks*, 16, n. 1, pp. 11-23.

Tanioka, R., Sugimoto, H., Yasuhara, Y., Ito, H., Osaka, K., & Zhao, Y. (2019), "Characteristics of Transactive Relationship Phenomena among Older adults, Care Workers as Intermediaries, and the Pepper Robot with Care Prevention Gymnastics Exercises", in *The Journal of Medical Investigation*, 66, pp. 46-49.

Tussyadiah, I., & Miller, G. (2019), "Nudged by a robot: Responses to agency and feedback", in *Annals of Tourism Research*, Elsevier, vol. 78(C), pp. 1-12.

Van De Maele, T., Verbelen, T., & Çatal, O. (2021), "Active vision for robot manipulators using the free energy principle", in *Front. Neurorobotics*, 15, 642780.

Van Wynsberghe, A. (2021) "Sustainable AI: AI for sustainability and the sustainability of AI", in *AI Ethic*s, 1, n. 3, pp. 213-218.

Wortham, R. H. (2020), "Transparency for Robots and Autonomous Systems: Fundamentals, technologies and applications", in *IET Control, Robotics and Sensors Series*, 130, Institution of Engineering and Technology, London.

Evaluating and measuring intelligence in Neural Language Models: a methodological approach

Francesco Bianchini
University of Bologna
francesco.bianchini@unibo.it

1. Introduction

In recent years, artificial intelligence (AI) systems have evolved at an increasingly rapid pace, encompassing multiple levels and perspectives. Alongside the natural technical advancements characteristic of this field, there has been significant progress in how AI systems and tools are accepted and integrated by users. As technology has advanced, users have developed a broader awareness of these tools. However, this awareness remains superficial and incomplete for many. For some users, this awareness manifests as a recognition of the existence of new AI tools. Others have come to appreciate their potential by using them for tasks of personal or professional interest. A smaller subset – typically more experienced users – has grappled with the actual limitations of these systems. This gradual acquisition of awareness, likely more widespread than at any previous stage in AI's history, has contributed to the broad diffusion of these systems[1]. These tools can be broadly categorized into two distinct yet partially overlapping groups: voice assistants and generative AI.

Voice assistants have emerged as tools rooted in decades of research in natural language processing[2]. They have significantly extended the horizon of interaction between humans and AI systems. Generative AIs, on the other hand, were developed with goals distinct from simply disseminating AI tools.

---

[1] Kelly *et al* (2023).
[2] Pieraccini (2012).

Francesco Bianchini, "Evaluating and measuring intelligence in Neural Language Models: a methodological approach", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 75-100

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

Voice assistants act as interfaces connecting the networked world with productive, commercial, and informational domains. They have also broadened access to these resources for many individuals, including those previously unable to read or write. By simplifying access to internet content through natural language, voice assistants enable easier interaction with the vast array of online resources. They achieve this by providing a user-friendly interface to access the network's content while collecting information from nearly anyone who interacts with them. Their primary goal is not to provide responses in an entirely human-like manner, but to accurately interpret *users' intentions* and give precise answers. In contrast, the focus of generative AIs lies in producing outputs – whether text, images, or multimedia content – that are not only relevant but also convincingly human-like. They are built to return a result that is as appropriate as possible from the point of view of interaction to be *cognitively* understood as human *by the user*. The purpose of a conversation with a generative language system is to create interactions that are indistinguishable from those with a human, both in originality and style. While both voice assistants and generative AIs rely heavily on language, their objectives diverge. Voice assistants openly function as AI tools, with their artificial nature visible to the user. Generative AIs, however, strive for an interaction so seamless that their artificiality fades entirely from the user's perception.

Therefore, compared to voice assistants, generative AI models dedicated to language present distinct characteristics that warrant closer examination, particularly in terms of the intelligence they exhibit. This article will explore Neural Large Language Models within the framework established by Turing (Section 2), address the challenges of evaluating and measuring the intelligence of AI systems in contemporary contexts also by formulating a new methodological approach (Section 3), and analyze generative AI models for language through this specific lens (Section 4). Finally, in the conclusion (Section 5), observations will be offered on the potential challenges and developments expected within this field in the near future.

2. Turing was right

Neural language models (NLMs) are artificial neural networks specifically designed for natural language processing (NLP) tasks. Among these, Large Language Models (LLMs) have gained prominence in recent years, representing a key area within generative AI. LLMs are built on the

Transformer architecture[3], which leverages an attention mechanism originally developed for machine translation, a foundational domain in NLP[4]. The Transformer introduces a self-attention mechanism[5], enabling the model to process text sequences by relating different positions within a sequence. Through iterative applications of self-attention, the model forms a holistic representation of the sequence. This approach enhances encoding and decoding processes, offering significantly faster performance compared to recurrent neural networks. Crucially, self-attention facilitates contextual understanding, allowing the model to represent a word's meaning dynamically based on the specific text or sequence in which it appears. As with other neural networks, these representations are vector-based, and computations occur through transformations across multiple intermediate layers.

The technical aspects of these models are essential for understanding their place within the broader category of statistical-predictive systems. This is why they have been described, in the context of language generation, as "stochastic parrots"[6]. Due to their capacity for generating conversational language, these models also potentially align with the concept of "thinking machines" as defined by Turing (1950) prior to the advent of AI. Turing envisioned machines capable of conversing with humans in a way that would make it indistinguishable whether they were interacting with a human or a specially programmed digital computer. As is well known – and extensively discussed in the literature on what is now called the Turing Test[7] – Turing did not specify the exact nature of such machines. He hypothesized they would likely need the ability to learn but offered no guidance on how these thinking programs should be constructed[8].

---

[3] Vaswani *et al.* (2017).
[4] Bahdanau *et al.* (2015).
[5] Vaswani *et al.* (2017).
[6] Bender *et al.* (2021).
[7] Cfr. Moor (2003).
[8] In fact, there is a significant gap between what Turing envisioned and contemporary LLMs, both in terms of objectives and underlying concepts. Turing's goal in his 1950 paper was to provide an operational means of addressing the question "Can machines think?" while avoiding philosophical entanglements. Linguistic interaction was one of the devices he employed to construct the hypothetical scenario, specifically to create a neutral ground for comparing human beings and appropriately programmed digital computers. This setup originated from the imitation game played between an interrogator on one hand, and a man and a woman on the other. Over time, however, and regardless of whether this was faithful to Turing's original intent, the focus on linguistic interaction became central. It eventually came to define the standard interpretation of the Turing Test, giving rise to a wide-ranging and productive debate.

These kinds of questions emerged a few years later with the advent and subsequent development of AI. However, Turing insists that such machines must be able to play the imitation game, regardless of the specific characteristics of the game itself: "it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man"[9]. If we consider the technical aspects of LLMs, they do not seem to align with what Turing had in mind. At best, they are learning machines, but the sense in which they "learn" is somewhat vague and does not easily lend itself to comparison with Turing's intended claim. Nevertheless, when we examine the actual functioning of pre-trained generative models based on Transformers, the scenario envisioned by Turing appears strikingly relevant. Natural language interaction with these models occurs through prompts – questions or suggestions posed to the program – where inputs can range from multiple examples (few-shot learning) to none at all (zero-shot learning). The progressive refinement of these systems yields natural language outputs that are largely indistinguishable from human-generated text or at least equally comprehensible. From this perspective, LLMs are already capable – and will likely become even more so – of passing the Turing Test in its classical form.

According to Turing (1950: 449), this outcome would not be surprising, given his prediction that within fifty years of his seminal article on computational machines and intelligence, a computer would be able to play and win the imitation game at least thirty percent of the time. In making this prediction, Turing does not concern himself with the specific characteristics of the system capable of winning the game, aside from a general reference to computational resources. What truly matters is that the interaction occurs in a human-like linguistic format. That said, not in all discussions on thinking machines did Turing disregard their cognitive characteristics[10]. However, his 1950 text focuses primarily on the possibility of natural language exchanges between humans and machines. This exchange serves a dual purpose: 1) it places both entities on neutral ground and 2) allows for discussions on any topic, functioning as a kind of generalist methodology. The latest generation of LLMs increasingly align with Turing's vision. They engage in *linguistic* exchanges, exhibit *generality* in the range of topics they can cover, and generate text that is *understandable* in a human-like manner. Even minimal interaction with the most advanced LLMs demonstrates that these

[9] Turing (1950: 437).

[10] Turing addressed the issue of the characteristics that a system capable of learning must have, for example, in a 1951 work, focusing on the role of memory and the way in which it can become increasingly complex and "human". See Turing (1951).

characteristics are met, making it clear that the Turing Test, in its original form, is easily passed.

The push toward simulating human thought is more evident in Turing's 1951 text, suggesting that he himself considered the Turing Test insufficient for assessing the presence of intelligence in a machine, at least when it comes to human-like intelligence: "my contention is that machines can be constructed which will *simulate* the behavior of the human mind very closely"[11]. Although this text is less frequently cited, it introduces a crucial concept for the later development of AI: simulation. It also reinforces the idea that the simulation of human intelligence was already a central theme for Turing. This perspective aligns with how we might evaluate neural network-based LLMs. While they can engage in human-like conversational interactions, they ultimately exhibit only verbal behavior. They are machines – programs – they do not understand the content of their own outputs. Instead, they generate coherent word sequences by computing probabilistic relationships between tokens. In essence, LLMs predict language with remarkable accuracy, but they do not embody intelligence in the sense we typically use the term. In some ways, they could be seen as a modern version of machines that elicit an Eliza effect – a phenomenon named after the ELIZA program developed by Weizenbaum (1966). This interpretation, however, does not fully capture Turing's vision[12]. And the key question remains: do these models actually exhibit intelligence?

An answer of this kind risks being too simplistic. LLMs undoubtedly exhibit a *form* of intelligence; after all, their outputs, whether text, code, or other content, are difficult to dismiss as unintelligent. The question of intelligence in LLMs has been already approached from multiple perspectives. For some scholars – for instance, Millière and Buckner (2024) – the concept of intelligence is too elusive to be meaningfully applied to LLMs. Others have examined the metaphors used to describe generative AI systems, particularly LLMs, to assess the implications of characterizing them as intelligent, especially from an anthropomorphic standpoint[13]. More recently, LLMs have also been employed to investigate various dimensions traditionally associated with intelligence, including different forms of understanding[14] and the relationship between LLMs and the brain[15].

[11] Turing (1951: 472, emphasis added). There is also a reference to simulation in Turing (1950), but only to compare the adult mind with the child mind.
[12] On imitation and LLMs see Boisseau (2024).
[13] Mitchell (2025).
[14] Miracchi & Titus (2024).
[15] Lamarre *et al.* (2022).

A more precise and relevant (for the aim of this article) question might be: can we meaningfully attribute intelligence to them? Or rather, when we speak of intelligence in reference to LLMs, what exactly are we referring to? This question arises precisely because their outputs compel us to recognize a form of intelligence; otherwise, we risk losing sight of what we consider cognitively valuable – content that can be used in epistemic contexts or at least remains intelligible within a cognitive framework. The challenge, then, is: how should we evaluate their intelligence?

3. The quest for evaluating AI intelligence

Since the earliest developments in AI, the challenge of evaluating intelligence in artificial systems has taken on a dual form. On one hand, it has followed in the footsteps of Turing and the Turing Test, generating numerous variations and fueling a decades-long debate[16]. On the other hand, various AI approaches have been examined to determine which best aligns with the goal of creating human-like or cognitively plausible intelligence, at least to some extent[17].This second line of inquiry, often intertwined with the evolution of cognitive science, rests on the assumption that artificially replicating human cognition is, by definition, a valid means of simulating intelligence. In other words, if intelligence is a defining characteristic of human beings, then reproducing their cognitive mechanisms, functions, and properties should lead directly to the simulation of intelligence. The main issue with this perspective is its excessive anthropocentrism. This concern has been particularly noted in relation to classical AI and its symbolic approach[18]. However, even in more recent developments in AI – shaped by the embodied turn in cognitive science and the rise of bio-inspired computational architectures – traces of anthropocentrism persist, in line with a view that underlines a partial continuity between classical approaches and new approaches to cognitive science[19].

Of course, the evaluation of intelligence of AI also forms part of a broader and long-standing debate concerning the nature of intelligence itself, a debate that has not always been approached from a human-centered perspective. In the context of AI, intelligence has, for instance, been investigated as a property of systems, often linked to rationality as a defining

---

[16] Moor (2003).
[17] Cfr. for example, Boden (2006).
[18] Preston (1991).
[19] Shapiro (2019).

feature of intelligent behavior[20]. More recently, scholars have explored the nature of both human and machine intelligence in relation to creativity[21], as well as to capacities such as perception, understanding, and abstraction within learning processes[22].

Recent developments have introduced new methods for evaluating intelligence in artificial systems, shifting the focus toward measuring intelligence rather than treating it as a simple yes/no question. These methodological approaches recognize that the issue is tied to a broader, unresolved question: what is intelligence?

Moreover, the challenge of attributing intelligence has gained increasing importance in recent years, driven by the widespread proliferation of AI systems. Over the past fifteen years, AI has transitioned from a specialized technological discipline, primarily confined to niche applications, to a widely accessible software technology used by the general public. The advent of LLMs has further accelerated this diffusion, leading to the growing, often unreflective, integration of AI systems into everyday life. This raises critical questions about how users perceive both the performance and the outputs of these systems. In particular, this development brings forth a range of ethical and societal concerns, spanning multiple domains, including culture, education, information dissemination and communication, marketing, and commerce, among others. The central thesis of this work is that the ways in which intelligence is attributed to and evaluated in AI systems are increasingly relevant for their appropriate deployment and integration into society. Furthermore, this form of evaluation is, at its core, an epistemic issue with significant epistemological dimensions. In the following pages, key guidelines are outlined for constructing such an evaluation framework, referring to Bianchini (2024) for a more detailed discussion.

The problem of attributing and evaluating intelligence goes deeper than the simple Eliza Effect mentioned earlier. It is not merely about the possibility of being “fooled” by systems that employ tricks to create the illusion of intelligence in their behavior or outputs. In other words, it is not just a contemporary manifestation of the broader human tendency to attribute intentionality or understanding as part of cognitive processing. This debate has been central to the philosophy of AI for decades[23] and remains active, particularly in discussions surrounding human-artificial system interaction,

[20] Russell (1997).
[21] Boden (2016).
[22] Mitchell (2019).
[23] Cfr. Dennett (1987) and Searle (1983).

especially in robotics[24]. However, the discourse on the attribution of intentional attitudes – while fundamental to the philosophy of mind and crucial in human-robot interaction – primarily concerns unreflective attribution. That is, it examines the natural human tendency to ascribe intentionality, and by extension, intelligence somehow, to non-human entities, particularly artificial systems, thereby granting them an appearance of cognitive/intelligent capacity. Recently, scholars have questioned whether such attribution occurs as widely as traditionally assumed. Some argue that certain forms of anthropomorphizing may be more myth than reality[25]. Nevertheless, even if such attributions are less pervasive than once believed, their persistence underscores the significance of this issue in AI research. It remains crucial not only for understanding human interactions with AI systems but also for assessing these systems both as practical tools and as subjects of theoretical analysis from the point of view of intelligence.

In this regard, it is necessary to take a further step and consider the interaction with AI systems, particularly in relation to evaluating their intelligence. As previously mentioned, the classic attribution of intentionality is largely considered an automatic cognitive act rather than a conscious assessment. A conscious attribution, however, is based on expected intelligence and can serve as the foundation for new approaches to measuring intelligence. The notion of *expected intelligence*, which is closely tied to an interactive approach to AI, provides a basis for evaluating the intelligence of an artificial system through the initial assumptions made by the user interacting with it. In this context, expected intelligence refers to the largely conscious tendency to engage with an artificial system from which an epistemically and/or cognitively relevant response is anticipated. In other words, expected intelligence functions as a precondition for recognizing, and thus evaluating, an artificial system as autonomous system. Without this consciously held precondition, the system's behaviors and outputs would not necessarily be interpreted within a meaningful framework and might instead be regarded as mere occurrences or mechanical reactions to specific stimuli. In AI, and particularly in fields such as robotics, the interactive approach relies on the concept of expected intelligence both to explain the behavior of artificial agents and to guide their design in relation to cognitively capable users, namely human beings. This concept thus serves as the starting point for evaluating attributed intelligence[26].

---

[24] Wykowska (2024).
[25] Coghlan (2024).
[26] Bacaro & Bianchini (2024).

The ability to assess the attributed intelligence – or lack thereof – of a system is crucial not only for understanding AI itself but also for evaluating its broader social and technological impact. This evaluation plays a significant role in addressing the Collingridge Dilemma[27], which highlights a fundamental challenge in technology governance: some technologies are difficult to predict in terms of their societal impact until they become widely adopted, yet by that time, they are often difficult to control or modify, particularly in terms of their standardized use. AI systems developed over the past decade fit this dilemma perfectly, especially those that are easily accessible and widely used. LLMs provide a clear example. Their rapid and widespread adoption is largely due to their impressive capabilities, yet their long-term impact remains under scrutiny. The widespread diffusion of LLMs, whose consequences remain difficult to fully anticipate, has given rise to a broad spectrum of ethical issues. These range from the potential amplification of misinformation and the reinforcement of biases to the economic and social impacts of their deployment, as well as concerns about reliability, particularly regarding the data on which these models are trained. While some of these challenges are common to all systems based on deep neural networks, they become especially critical in domains where text production and the use of knowledge are foundational, such as education, or where data usage, transparency, and reliability are essential prerequisites for application, as in the medical field[28]. Developing a conscious evaluation of the expected intelligence of such systems – beyond merely assessing their efficiency and accuracy – could offer a means of mitigating the challenges posed by the Collingridge Dilemma, particularly where predictive limitations arise, and the related ethical issues.

Let us now examine in more detail how to evaluate the expected intelligence of an AI system. In the first place, this issue can be seen as equivalent to measuring the intelligence of an artificial system deemed intelligent. The systematic analysis of intelligence measurement in AI has gained attention only relatively recently and has led to two primary characterizations[29]: a) intelligence as a set of task-specific skills; b) intelligence as a general ability to learn and perform open-endedly. In the first case, the focus is on measuring the accuracy of an AI system's performance. Here, no generalization occurs – neither within the system itself (narrow generalization) nor through developer-implemented methods (broad

[27] Collingridge (1980).
[28] Ong *et al.* (2024).
[29] Hernández-Orallo (2017), Chollet (2019).

generalization). In the second case, the aim is to assess how well specific abilities can be generalized across multiple domains. This approach is reminiscent of Newell, Shaw, and Simon's (1959) General Problem Solver and is further developed in modern cognitive architectures such as SOAR and ACT-R.

The first approach – measuring task-specific performance – appears particularly well-suited for evaluating AI systems. This is because it allows for the construction of a measurable value scale, typically based on accuracy. Such measurements are often carried out relative to a predefined standard or as an average over multiple performances. In this framework, assessing AI intelligence usually entails evaluating task-oriented performance on a scale, where a "good" or "poor" performance is determined by specific parameters. This process is inherently deliberate and guided by a well-defined objective. Hernández-Orallo (2017) identified three types of methods and metrics aligned with this perspective, focusing on: 1) human discrimination; 2) problem benchmarks; 3) peer comparison. The first approach is inherently subjective and remains within an anthropocentric framework. The other two involve comparison either with a predefined standard or with the average performance of other systems or human participants performing the same task. In this sense, they can be considered more objective and provide effective parameterization, even if they are limited to highly specific tasks, such as categorization in a neural network or user preference profiling.

The challenge arises with generality – specifically, the evaluation of AI systems' intelligence across different domains and from an indeterminate perspective. In particular, how can we assess intelligence based on abilities, focusing on broader cognitive aspects? The risk here is falling into anthropocentrism, searching for cognitive traits within AI systems. While this approach aligns with cognitive science's historical research programs[30], it differs from evaluating a system's expected intelligence, an issue that remains neutral regarding whether AI systems possess cognitive qualities. On the other hand, adopting a neutral formal standard for evaluating AI intelligence – such as one based on algorithmic information theory[31] – risks resulting in an opaque assessment. This is because objective measurement elements would primarily relate to different dimensions of algorithmic complexity. However, complexity and information are not *directly* equivalent to intelligence. In other words, while intelligence can be seen as a property of

[30] Boden (2006).
[31] Chaitin (1987).

complex systems, it does not follow that every complex system capable of processing information is necessarily intelligent.

To address the challenge posed by generalist approaches to AI – particularly in assessing their adaptability across multiple contexts, a key hallmark of intelligence – three distinct theoretical responses can be considered. First, one might argue that AI systems are not intelligent at all but merely instruments of action[32]. This perspective rests on the assumption that an intelligent outcome is not always an intelligent behavior, or the result of an intelligent behavior. While this "eliminativist" stance on AI intelligence may seem too radical, it has the merit of distinguishing between intelligence as an intrinsic property of the system and the attribution of intelligence to the system itself.

A second possible response focuses on the social and interactive aspects of AI systems[33]. The study of human-AI interaction has a long history, and interactional perspectives have gained increasing relevance, partly due to the rise of embodied approaches, such as enactivism, within cognitive science, particularly in relation to artificial systems. Without committing to a specific theory of cognition, a general assumption in this view is that, in most cases involving human users and AI systems, the attribution of intelligence by the human user, often in real-time, is crucial for achieving optimal results and effective interaction. In other words, without the presumption of a shared cognitive framework, which falls within the broader concept of intelligent behavior, meaningful interaction becomes unlikely. Instead, the AI system risks being reduced to a mere tool used by the human operator.

Finally, a third possible response arises from the debate on the attribution of mental states to artificial systems[34]. The tendency of humans to ascribe mental states –particularly to robotic artifacts – is one possible explanation for the way we interact with certain AI systems. This attribution is not necessarily limited to robotic systems; it can also extend to other artificial entities perceived as intelligent. Within this perspective, the debate remains open regarding the ontological status of these attributed mental states and the various approaches to assigning intentionality to AI systems[35]. Nevertheless, while attributing mental states can serve an explanatory role in understanding human-AI interaction, it does not necessarily address the issue of intelligence itself. Intelligence, in this sense, remains conceptually distinct from the cognitive elements we might identify when evaluating these systems. In other

[32] Floridi (2023).
[33] Cristianini *et al.* (2023).
[34] Thelmann *et al.* (2022).
[35] Larghi & Datteri (2024).

words, the attribution of intelligence to an AI system appears to be independent of what we believe is happening within the system, even from an attributional standpoint.

If, on the one hand, we wish to avoid overly deflationary positions regarding the intelligence of AI systems, and, on the other, set aside considerations about how these systems are designed or aligned with recognized cognitive systems – primarily humans – the behavioral perspective remains the most central approach for the conscious attribution of intelligence[36]. This perspective, which can be seen as partly inheriting Turing's legacy, allows us to analyze the attribution of intelligence from the user's standpoint, emphasizing its role as an essential requirement for the epistemic, applied, and ethical functioning of AI systems.

The attribution of intelligence from the user's perspective can be developed along four dimensions[37]:

Before interaction – Based on the user's preliminary knowledge of the AI system.

During interaction – While actively using or engaging with the system.

Post-interaction – Evaluating the system's performance and the outcomes it produces.

Over repeated interactions – Assessing intelligence attribution over time, considering potential variability in perception.

In all these cases, the system's behavior is evaluated in broad terms. This evaluation can concern both performance on a specific task, especially when repeated with varying results, and the system's behavior from a more general perspective. The latter involves determining whether the system demonstrates general capabilities beyond isolated tasks, indicating a broader implementation of intelligence.

Finally, different metrics can be devised to best capture the four dimensions of evaluation, depending on the specific context. Without aiming for exhaustiveness, at least two broad categories of applicable metrics can be identified.

The first category includes metric formats based on scalar dimensions within a defined range: for example, Likert-type scales. These can vary in granularity depending on the level of detail desired (e.g., five-point, seven-point, or ten-point scales). A higher level of detail may be appropriate for assessing the attribution of intelligence in scientific or experimental settings,

---

[36] For a behavioral perspective on evaluation in terms of prediction see Cevolini, Esposito (2022).

[37] For a more detailed description see Bianchini (2024).

while lower-resolution scales can support self-assessment by users regarding their interaction with an AI system. In such cases, the aim may be to promote user self-awareness and responsibility, to implement nudging strategies, or to generate aggregate rating data that can inform system design or incremental improvements. It is also worth noting that the four dimensions allow for a temporal assessment of the attribution of intelligence within the interactive process, whether it is increasing, decreasing, or remaining stable. This temporal perspective can help identify the phases of interaction in which perceived or attributed intelligence is heightened or diminished. For instance, a decreasing attribution over time may indicate that the system is perceived as displaying a weak degree of "artificial intelligence", and thus as being less reliable or accurate in relation to user expectations.

The second type of metric could instead leverage the direct relationship with the user, considered as a median point of reference. From this midpoint, the user would assign scores indicating whether the intelligence attributed to themselves is greater or lesser than that attributed to the system at various stages of the interaction. As with the first type, these metrics could vary in granularity depending on their intended purpose. The goal of this approach is to place the user even more centrally in the process of attribution, encouraging them to assess their own intelligence in comparison to that of the system. This can have several theoretical implications for research on human-AI interaction, as well as practical benefits. For example, it may promote more conscious and constructive use of the system, help identify weaknesses in the interaction, and foster more responsible usage, especially in contexts where there is a risk of user deskilling (among which the educational one).

4. Measuring expected artificial intelligence and the case of LLMs

Beyond the potential metrics that could be developed using these four dimensions – aimed at refining the measurement of AI intelligence across different application domains – this proposal seeks to capture a fundamental practical principle: *intelligence is attributed when it is expected, and it is expected when it is attributed*. This principle applies particularly to AI systems, which are defined within the broader field of artificial *intelligence* and are characterized by their capacity to implement intelligent behavior autonomously, another key criterion of AI.

This discussion has significant methodological implications for investigating AI in relation to human intelligence and cognition. Since the inception of AI, researchers have explored the possibility of constructing AI systems as a means of understanding human cognition and its processes[38]. However, the principles underlying this undertaking can be generalized. The assumption underlying our approach is that expected intelligence – attributed to a system by human observers – is coupled with *something* underlying that enables intelligent behavior. This something, in turn, serves as a preliminary condition for recognizing intelligence. Such an assumption carries two important implications. First, it justifies treating the system as intelligent, meaning we must expect it to perform actions we consider intelligent; otherwise, we risk falling into deception or misconception. Second, it places a demand on human intelligence itself: the system's behavior must be authentically intelligent, rather than a collection of superficial tricks that undermine the legitimacy of considering it truly intelligent.

The crucial question, then, is: where do we draw the line between authentic intelligence and mere imitation? To avoid anthropomorphism or the assumption that intelligence must emerge from specific internal mechanisms modeled on human cognition, we can turn to the concept of expected intelligence as a measurable phenomenon. This allows us to address the boundary between intelligence and non-intelligence in a more gradual and pragmatic way, aligned with real-world interactions between humans and AI systems. This behavioral perspective has the further advantage of avoiding a human-centric commitment to what constitutes intelligence. In other words, the processes that give rise to intelligence in an AI system do not necessarily have to mirror those found in human cognition.

On the other hand, this perspective carries the risk of leading to an overly anarchic approach to the attribution of intelligence. If intelligence could be ascribed to virtually anything without a clear justification, the concept itself might lose its meaning. Therefore, it seems necessary to also consider the issue from the opposite standpoint. To avoid such conceptual chaos – where intelligence could be arbitrarily attributed without a solid basis – there must be some criterion to justify the attribution. This criterion could take the form of a mechanism, a technique, a dynamic interaction, a mathematical or statistical function, or any other structured method. While this criterion does not necessarily need to be predetermined – allowing for a certain degree of flexibility or a standby approach – it must still exist in some form to preserve the coherence of the notion of artificial intelligence as applied to the system in question. In practice, the loss of this notion is not what we observe in the real world. Instead, the attribution of intelligence, at least to some degree, to AI systems is something we do continuously.

[38] Cordeschi (2002).

Leaving aside analytical metrics, let's attempt to transform the four dimensions by which we define the evaluation and measurement of an AI system's intelligence into a methodological approach. It will then be considered its epistemological significance. The steps of this methodological process could be as follows:

1. *Assuming* the possibility of using or interacting with an AI system.
2. *Expecting* intelligence in the system.
3. *Attributing* intelligence (hypothetically) to the system.
4. *Attributing* or "*finding*" intelligence (actually) to the system.
5. *Identifying* the "reason" of intelligence in the AI system (the research-oriented step).

The first four steps can be applied whenever one encounters an AI system or a system presented as such. Confirming step 4 in this process amounts to recognizing the system as genuinely intelligent and potentially assigning a measurement index to this characteristic. Step 5 is optional and relevant primarily when situating the system within a particular AI framework, or multiple convergent AI approaches, for research, regulatory, ethical, or legal purposes.

In more detail, the four dimensions previously described can be mapped onto this methodological process as follows. The *"before"* dimension corresponds to steps 1-2-3, as it involves moving from the initial assumption to a hypothetical attribution of intelligence. The *"during"* dimension spans steps 2-3-4, since it covers the transition from the evaluation of expected intelligence to its actual attribution to the system. The *"after"* dimension pertains to steps 3-4 and specifically involves the ex-post assessment of the shift from hypothetical to actual attribution. Finally, the *iterative* dimension encompasses steps 1–2–3–4, as the evaluation process is designed to be repeated over time.

In general, steps 1 to 5 can be understood as a form of reverse engineering through interaction. That is, rather than beginning with the acknowledgement of predefined cognitive capabilities, one could start from direct engagement with the system. More specialized competencies of experts would come into play at a later stage of analysis. This approach would enable even non-experts to engage with AI systems in an informed manner, using an initial heuristic evaluation and measurement method to navigate their interactions effectively. It is important to note that as AI systems become increasingly integrated into daily life and accessible to all types of users, this methodological framework will be crucial. It will serve not only as a means of maintaining oversight of

AI systems but also as a foundation for sustainable and informed interactions with them – an essential aspect of the society of the coming decades.

The applicability of this method is broad within the field of AI and extends to all users of AI systems, including those interacting with content profiling tools, voice assistants, medical and educational technologies, autonomous vehicles, and even autonomous weapons. In all these cases, both general users and experts – though not necessarily AI specialists – can engage with AI systems and analyze their interactions from an intelligence-based perspective.

Among the most prominent AI systems today are generative AI systems, particularly neural large language models, already mentioned in the earlier sections. LLMs possess distinctive characteristics that make them especially well-suited for evaluation through the methodological framework outlined above. Their performance can be assessed in a task-oriented manner across various domains, and they belong to the broader neural network paradigm, which is explicitly designed to handle diverse tasks. This inherent generality, however, presents a challenge: it is often too expansive to be meaningfully evaluated as a single entity. Nevertheless, LLMs demonstrate linguistic competence across a vast array of subjects, suggesting that language might represent the appropriate level of generality at which to assess their intelligence. Moreover, many contemporary models are multimodal, capable of processing text, images, and code as inputs. Essentially, these systems perform a specific task, language processing, but in a way that connects to a wide range of topics. In this sense, they can be seen as task-specific systems exhibiting a form of general intelligence – namely, linguistic intelligence in the broad sense. For this reason, LLMs appear to strike the right balance between specialization and generality for assessing AI intelligence: they are neither so narrowly focused as to reduce their cognitive potential to a single capability nor so broadly defined as to make their intelligence indistinguishable from mere computational complexity.

From the perspective of LLMs, the five steps introduced are easily applicable. The growing confidence in these systems parallels their rapid diffusion, which aligns perfectly with Collingridge's dilemma[39]. This, in turn, underscores the need for a more conscious and responsible use of "intelligent" tools. Let us now explore how, at an epistemological level, these AI systems can be evaluated in relation to textual production.

First, there is now broad consensus that LLMs should be regarded as intelligent tools, not merely in the generic sense of being AI systems, but in

[39] Collingridge (1980).

the more substantive sense of enabling the production of outputs recognized as intelligent. This directly leads to step 2: the expectation that the system will generate texts that are coherent, meaningful, relevant to user queries, and cognitively adequate for human understanding. This step – akin to a "Turing step" – is generally satisfied, particularly by the most advanced LLMs, which can respond to a vast range of natural language requests across an indeterminate number of topics. Step 3 follows: the hypothetical attribution of intelligence to the system itself, rather than just its outputs. This step is a generalization, where the system's intelligence is empirically inferred from the quality of its textual productions and extrapolated into a broader hypothesis of general intelligence. Step 4 involves confirming this attribution of intelligence, which can be assessed using the four temporal dimensions of interaction previously mentioned. These dimensions can also be quantified to allow for a more gradual evaluation of intelligence, moving beyond a binary distinction between intelligence and non-intelligence. For instance, intelligence can be evaluated through a metric that assesses the comprehensibility, relevance, and coherence of the generated texts, features typically associated with intelligence. Similarly, the fourth dimension, repeated use, can help determine the system's reliability: whether it produces false information, when it starts generating hallucinations (i.e., plausible but incorrect content based on its training data, now a well-documented characteristic of LLMs[40]), and the extent to which it exhibits standardization or stylistic repetition. This longitudinal evaluation can also assess whether errors are present and how they evolve over time.

The value of steps 1–4 lies in their ability to provide all users with a framework for evaluating AI systems, fostering a bottom-up approach that enables meaningful interaction with intelligent systems. This evaluation allows users to assess the system's potential actions, activities, or behaviors from the perspective of intelligent understanding. Such an approach not only helps in interpreting the capabilities and limitations of LLMs but also extends to other AI systems. Consider, for example, interactions with fully autonomous vehicles, AI-driven medical applications, or even autonomous weapons. A precise understanding of their "intelligent" behavior is essential for ensuring safe and effective interaction, especially in high-stakes scenarios where errors could lead to disastrous and irreparable consequences.

The application of steps 1–4 can yield particularly interesting and practical outcomes for users not primarily concerned with research purposes, especially in the case of LLMs. While the attribution of intelligence in other

[40] See Farquhar *et al.* (2024).

AI systems often serves as a prerequisite for assessing their reliability, LLMs introduce a distinct epistemic dimension. Consider, for instance, navigational systems or self-driving vehicles. We trust their intelligence insofar as we delegate to them tasks that we would typically perform using our own cognitive abilities. This trust largely hinges on the extent to which we attribute intelligence to these systems, especially since, in most cases, we lack detailed knowledge of the technical mechanisms underlying their autonomous operation. Take the extreme example of a monorail transporting passengers between terminals in an airport without a human operator. It is relatively easy to trust such a system because we can readily imagine the limited and well-defined nature of the task, which seems to require only a modest level of intelligence, if any, by everyday standards. We can roughly grasp how its autonomy functions and feel comfortable attributing it with just enough intelligence to fulfill that role. By contrast, navigation systems or autonomous vehicles involve a far greater number of variables, and their functioning depends on mechanisms that are more opaque and harder to conceptualize. In these cases, the attribution of intelligence is closely tied to the reliability we are prepared to grant them, perhaps based on direct experience or observed behavior. A mistake or failure would diminish our trust, effectively lowering the degree of intelligence we attribute to the system. As a result, we may become reluctant to rely on it again unless significant improvements and verifiable changes are made.

Let us now consider the case of LLMs. In this context, we cannot merely observe their behavior as with other AI systems; rather, we must assess the products they generate through interaction to judge their intelligence. Unlike in other systems, intelligence here is not conflated with reliability, something we may be willing to compromise on, as long as we are aware of it and the system remains useful, but is instead tied to usability itself. In the case of LLMs, the pragmatic dimension gives way to a cognitive-epistemic one. If we did not regard LLMs as intelligent, that is, as capable of producing coherent, comprehensible texts aligned with our prompts and responsive to the real world, we would have no reason to use them. To attribute such capabilities, however, we must expect LLMs to produce "intelligent" texts: texts that possess semantic interpretability, epistemic content, and – crucially – some trace of the evidentiary or inferential structure that would allow us to confirm or contest their claims. Only by attributing a degree of intelligence to an LLM can we evaluate its textual outputs according to these criteria, much as we routinely do with human interlocutors. If an LLM fails to meet these standards, we cease to use it. If it succeeds, then, even if only with regard to its outputs, we are implicitly attributing to it a minimal cognitive common ground. This common ground may shift depending on context, users, or over time. However, the more robust and recognizable this shared cognitive basis becomes, the more intelligence we attribute to the model, and the more inclined we are to engage with it. Importantly, this attribution does not require that the LLM possess intelligence of the same kind as human beings. Even outputs containing hallucinations may offer useful information, despite their misleading nature. We recognize the value in such texts because we attribute to the system a degree of intelligence, albeit a limited one, sufficient to distinguish them from mere juxtapositions of words devoid of meaning or relevance.

The key point with LLMs is that, unlike other AI systems, their adoption has been significantly more widespread and rapid. Moreover, unlike other forms of AI, it is difficult to define a fixed set of instructions to learn how to use them correctly. Instead, it is through use and interaction that users gradually learn how to operate them effectively. For this reason, attributing intelligence to these systems becomes a necessary first step, one that users must continually take to engage with them appropriately. This consideration also applies to domain experts who may not be directly involved in AI research. For instance, professionals such as lawyers or physicians can rely on LLMs to support their work, but they must be able to assess the degree of intelligence these systems display in their respective fields. This is essential to avoid risks such as bias or epistemic injustice. In such cases, knowing how the systems work is not sufficient. Proper use of these tools depends on the textual knowledge they produce in interaction, more specifically, on the user's ability to interpret their output appropriately and to formulate prompts competently, in line with the capabilities attributed to the system.

Step 5 of this methodological approach addresses more advanced research interests and involves experts working with AI in various capacities. The question of what underlies the intelligence observed in these systems is both a matter of practical design and implementation and a theoretical issue within an epistemological framework. Thus, answering the question "what is intelligence due to in this reverse engineering process?" has both practical and conceptual implications. For instance, determining whether intelligence in the system arises from statistical-predictive methods, mechanisms, network topology, structural design, architecture, inferential and/or representational abilities, or a combination of these factors can provide insights in multiple ways. It can inform the development of more efficient AI systems, enhance our understanding of intelligence and cognition, and help explain why AI systems are often perceived as intelligent from different perspectives.

An analytical consideration of the "reason" behind intelligence in AI systems can contribute to addressing several key challenges. It can aid in solving the problem of AI explainability[41]; it can help overcome anthropomorphism in the analysis of AI systems by identifying techniques that, while distinct from human reasoning, are nonetheless effective within specific programming domains, such as certain cases of supervised learning[42]; it can tackle semantic issues like symbol grounding, which remain relevant even in the latest AI systems, particularly in neural LLMs[43]. Additionally, this approach can support the development of models capable of inferring or deriving others' intentions and beliefs, so provided with a form of Theory of Mind[44]. Finally, and perhaps most importantly in relation to LLMs, such a methodological perspective can help determine the epistemic reliability of the texts these models generate. Specifically, it can assess to what extent we can trust the knowledge embedded in their outputs, both in particular cases and in general, thereby allowing us to evaluate their strengths and weaknesses as "epistemic authorities"[45].

A final consideration must be given to the risk of anthropomorphism, which increasingly concerns AI systems, especially generative ones, such as large language models (LLMs). Since the four evaluation dimensions outlined above, along with the proposed methodological process, are grounded in interaction between the AI system and the human user, and since the evaluation is carried out by the user on the basis of that interaction, the risk of anthropomorphic attribution is heightened. In other words, there is a growing tendency not only to interpret the behaviors and outputs of the system as anthropomorphic, but also to expect exclusively such behaviors, thereby selecting or misinterpreting those that fall outside this frame. This risk, however, is inherent in any process involving the attribution of cognitive features. The evaluation methodology proposed here should thus be understood in continuity with broader philosophical reflections on intentionality. Dennett himself – one of the most influential theorists of intentionality – argued that a necessary condition for attributing intentional states to a system is the presence of rationality, a rationality modeled on the human mind and shaped by evolutionary processes[46].

---

[41] Miller (2023).
[42] Watson (2019).
[43] Pavlick (2023).
[44] Nguyen & Gonzalez (2022).
[45] Ferrario *et al.* (2024).
[46] Dennett (1987, 1991).

The attribution of intelligence can be understood as a renewed form of attributing intentionality, one that focuses more on behavior and outputs than on the internal states of a system. However, the criteria used for such evaluation risk falling into the same anthropomorphic assumptions. How can we judge something to be intelligent except on the basis of what we already consider to be intelligent? Admittedly, knowing that we are dealing with an artificial system should prompt us to suspend judgment regarding the forms of intelligence we attribute, considering them with broader scope and a greater openness to possibilities beyond those supported by the "reasons" discussed in step 5 (e.g., similarity in structure or mechanisms with human beings). Yet even in this broader framework, the risk remains. If we combine the tendency toward anthropomorphism with automation bias, that is, the human predisposition to favor the outputs of artificial systems[47], we may similarly overestimate or over-rely on these systems' cognitive capacities. Just as automation bias can lead to an undue acceptance of machine-generated suggestions in decision-making, it can also foster, by analogy, an inflated attribution of intelligence to these systems. This risk becomes particularly pronounced when the system engages in human-like interaction, as in the case of linguistic exchanges with LLMs. This recurring challenge in the development of AI systems may be mitigated by cultivating greater user awareness and responsibility. As AI systems are continuously modified and improved, their performance becomes increasingly difficult to distinguish from human-like behavior, blurring the boundaries and increasing the likelihood of misattribution. One of the key aims of the four interactive dimensions proposed for evaluating the attribution of intelligence is precisely to foster this kind of awareness.

5. Conclusion

This paper aimed to address the evaluation of AI systems within the domain of neural network-based LLMs. The discussion of these models' intelligence began with an analysis of Turing's ideas, updated in light of the capabilities and behaviors of LLMs. The evaluation of intelligence was then reframed beyond the mere detection of its presence or absence in AI systems, particularly in LLMs. Recent developments in the debate on AI intelligence measurement were examined, highlighting the current focus on two main approaches: the evaluation of task-oriented systems, which excel in specific

[47] Skitka *et al.* (1999).

domains, and abilities-oriented systems, which demonstrate a broader and more general form of intelligence.

The proposal presented in this paper emerges from a reversal of perspective. Given the widespread adoption of AI systems, these technologies can now be considered accessible to a large number of users. Starting from the users and their *interactions* with AI, this paper argues for the necessity of fostering a conscious and informed use of these systems. This begins with examining the intelligence attributed to AI in relation to the intelligence expected within the user's knowledge context. Such conscious use can be guided by measuring the characteristics of AI intelligence along four temporal dimensions of interaction: before, during, after, and iterative. These dimensions can be translated into metrics to evaluate the various stages of the methodological approach proposed in the second part of this paper. This approach redefines the attribution and identification of intelligence by prioritizing user interaction over the intrinsic nature or design of the system. In this framework, understanding the underlying mechanisms of AI becomes the final step rather than the starting point, offering benefits across various fields while minimizing anthropocentrism and anthropomorphism. Neural LLMs serve as a prime example of widely adopted, interactive AI systems capable of generating behavior commonly perceived as intelligent. The proposed methodological approach has been applied to these models to illustrate potential research directions on LLMs and to explore the nature of intelligence in artificial systems.

References

Bacaro, M., & Bianchini, F. (2024), "Artificial Intelligence as Expected Intelligence", in F. Bianchini, V. Fano, P. & Graziani (eds.), *Current Topics in Logic and the Philosophy of Science. Papers from SILFS 2022 postgraduate conference*, College Publications, Rickmansworth, pp. 89-115.

Bahdanau, D., Cho, K., & Bengio, Y. (2015), "Neural machine translation by jointly learning to align and translate", conference paper in International Conference of Learning Representations 2015.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021), "On the dangers of stochastic parrots: can language models be Too big?", in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bianchini, F. (2024), "Evaluating Intelligence and Knowledge in Large Language Models", in *Topoi*, pp. 1-11. https://doi.org/10.1007/s11245-024-10072-5

Boden, M. (2006), *Mind as Machine. A History of Cognitive Science*. Oxford, Clarendon Press.

Boden, M. (2016). *AI. Its Nature and Future*. Oxford, Oxford University Press.

Boisseau, É. (2024), "Imitation and Large Language Models", in *Minds and Machines*, 34, 42. https://doi.org/10.1007/s11023-024-09698-6

Cevolini, A., & Esposito, E. (2022), "From Actuarial to Behavioural Valuation. The Impact of Telematics on Motor Insurance", in *Valuation Studies* 9 (1), pp. 109-39. https://doi.org/10.3384/VS.2001-5992.2022.9.1.109-139.

Chaitin, G. J. (1987), *Algorithmic information theory*, Cambridge, Cambridge University Press.

Chollet, F. (2019), "On the Measure of Intelligence", in *arXiv*:1911.01547v2, https://doi.org/10.48550/arXiv.1911.01547

Coghlan, S. (2024), "Anthropomorphizing Machines: Reality or Popular Myth?", in *Minds and Machines*, 34, pp. 1-25. https://doi.org/10.1007/s11023-024-09686-w

Collingridge, D. (1980), *The Social Control of Technology*, New York, St. Martin's Press.

Cordeschi, R. (2002), *The discovery of the Artificial. Behavior, Mind and Machines before and beyond Cybernetics.* Dordrecht/Boston/London: Kluwer Academic Publishers.

Cristianini, N., Scantamburlo, T., & Ladyman, J. (2023), "The social turn of artificial intelligence", in *AI & Society*, 38, pp. 89–96. https://doi.org/10.1007/s00146-021-01289-8

Dennett, D. C. (1987), *The Intentional Stance*, Cambridge Mass., The MIT Press.

Dennett, D.C. (1991), *Consciousness Explained*, Boston, Little, Brown and Co.

Farquhar, S., Kossen, J., Kuhn, L. *et al.* (2024), "Detecting hallucinations in large language models using semantic entropy", in *Nature*, 630, pp. 625–630. https://doi.org/10.1038/s41586-024-07421-0

Ferrario, A., Facchini, A., & Termine, A. (2024), "Experts or Authorities? The Strange Case of the Presumed Epistemic Superiority of Artificial Intelligence Systems", in *Minds and Machines*, 34, 30. https://doi.org/10.1007/s11023-024-09681-1

Floridi, L. (2023), *The Ethics of Artificial Intelligence. Principles, Challenges, and Opportunities*, Oxford, Oxford University Press.

Hernández-Orallo, J (2017), "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement", in *Artificial Intelligence Review*, 48, pp. 397-447. https://doi.org/10.1007/s10462-016-9505-7

Kelly, S., Kaye, S., & Oviedo-Trespalacios, O. (2023), "What factors contribute to the acceptance of artificial intelligence? A systematic review", in *Telematics and Informatics*, 77, https://doi.org/10.1016/j.tele.2022.101925.

Lamarre, M., Chen, C., & Deniz, F. (2022), "Attention weights accurately predict language representations in the brain", in *bioRxiv*, 2022-12.

Larghi, S., & Datteri, E. (2024), "Mentalistic Stances Towards AI Systems: Beyond the Intentional Stance", in A. Aldini (eds), *Software Engineering and Formal Methods. SEFM 2023 Collocated Workshops. SEFM 2023 Lecture Notes in Computer Science*, 14568. Cham, Springer. https://doi.org/10.1007/978-3-031-66021-4_2

Millière, R., & Buckner, C. (2024), "A Philosophical Introduction to Language Models, Part II: The Way Forward", in arXiv: 2405.03207

Mitchell, M. (2019), *Artificial Intelligence: A Guide for Thinking Humans*, New York, Farrar, Straus and Giroux (FSG).

Mitchell, M. (2025), "The metaphors of artificial intelligence", in *Science*, 386, 6723, DOI: 10.1126/science.adt6140

Miller, T. (2023), "Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI", in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

Miracchi Titus, L. (2024), "Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy", in *Cognitive Systems Research*, 101174.

Moor, J. H. (ed.) (2003), *The Turing Test. The Elusive Standard of Artificial Intelligence*, Dordrecht, Springer. https://doi.org/10.1007/978-94-010-0105-2

Newell, A., Shaw, J. C., Simon, H. A. (1959), "Report on a general problem-solving program", in *Proceedings of the International Conference on Information Processing*, pp. 256–264.

Nguyen, T. N., & Gonzalez, C. (2022), "Theory of Mind from Observation in Cognitive Models and Humans", in *Topics in Cognitive Science*, 14, pp. 665–686. 10.1111/tops.12553

Ong, J. C. L., Chang S. Y., Wasswa, W., Atul, J. B., Nigam, H. S., Lita, S. T. C. *et al.* (2024), "Ethical and regulatory challenges of large language models in medicine", in *The Lancet Digital Health*, 6, 6, e428 - e432. 10.1016/S2589-7500(24)00061-X External Link

Pieraccini, R. (2012), *The Voice in the Machine. Building Computers That Understand Speech*, Cambridge Mass., The MIT Press.

Preston, B. (1991), "AI, anthropocentrism, and the evolution of 'intelligence'", in *Minds and Machines*, 1, pp. 259–277. https://doi.org/10.1007/BF00351181

Russell, S. (1997), Rationality and Intelligence, in *Artificial Intelligence*, 94, pp. 57-77.

Searle, J. (1983), *Intentionality: An Essay in the Philosophy of Mind*, New York, Cambridge University Press.

Shapiro, L. (2019), *Embodied cognition*, 2nd edition, New York, Routledge.

Skitka l. j., Mosier, K. L., & Burdick, M. (1999), "Does automation bias decision-making?", in *International Journal of Human-Computer Studies*, 51, 5, pp. 991-1006, https://doi.org/10.1006/ijhc.1999.0252

Thellman, S., de Graaf, M., & Ziemke, T. (2022), "Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings", in *Journal of Human-Robot Interaction*, 11(4). https://doi.org/10.1145/3526112

Turing, A. M. (1950), "Computing Machinery and Intelligence", in *Mind*, 59, pp. 433–460, (reprinted in J. Copeland (ed.), *The essential Turing,* Oxford, Oxford University Press, 2004, pp. 441–464).

Turing, A. M. (1951), "Intelligent Machinery. A Heretical Theory", in J. Copeland (ed.), *The essential Turing,* Oxford, Oxford University Press, 2004, pp. 472–475.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017) "Attention is all you need", in *Proceedings of 31st International Conference on Neural Information Processing Systems* (NeurIPS 2017), Curran Associates, Red Hook, NY, pp. 6000-6010.

Pavlick, E. (2023), "Symbols and grounding in large language models", in *Philosophical Transactions of Royal Society A*, 381, 20220041. https://doi.org/10.1098/rsta.2022.0041

Watson, D. (2019), "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence", in *Minds and Machines*, 29, pp. 417–440. https://doi.org/10.1007/s11023-019-09506-6

Weizenbaum, J. (1966), "ELIZA. A computer program for the Study of natural language communication between man and machine", in *Communications of the ACM*, 9, pp. 36-45. doi:10.1145/365153.365168

Wykowska, A. (2024), *Intentional Stance Towards Humanoid Robots. Lessons Learned from Studies in Human-Robot Interaction*, Cham, Springer. https://doi.org/10.1007/978-3-031-65483-1

Umwelt and cities
Explanatory and Pragmatic Usefulness

Stefano Carlini
Scuola Normale Superiore
stefano.carlini@sns.it

Introduction

When we think about a city, we often picture things like buildings, vehicle traffic, the bustle of people in the streets etc. Less frequently, however, we imagine the huge number of non-human animals that inhabit the urban context with us. Even more rarely do we think of how these living beings may experience the city, by interpreting the environment in ways that can be very different from our own.

In this paper, I argue for the need to integrate the field of urban ecology with the ways in which animals perceive and interpret the elements of the cities. This operation, I claim, helps both in explaining some behaviours of urban fauna and in managing it. To do this, I propose to use the concept of *Umwelt*, first developed by Jakob von Uexküll at the beginning of the 20th century. This notion, while still struggling to take its place in urban ecology, is now gaining more and more attention, especially in the cognitive sciences (Feiten, 2020).

I will begin by showing the main traits of Uexküll's *Umwelt* theory (§1). Then, I will present the two main interpretations of *Umwelt* in the cognitive sciences literature (§2), and I will apply them to some phenomena of urban ecology, showing their usefulness both in explaining urban-animal's behaviours (§3) and in designing urban fauna management strategies (§4).

Stefano Carlini, "*Umwelt* and cities. Explanatory and Pragmatic Usefulness", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 101-121

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

1. Von Uexküll's *Umwelt* Theory

Developed at the beginning of the 20th century by the biologist Jakob von Uexküll (1928; 1934; 1982), the concept of *Umwelt* – "world (*Welt*) around (*Um*)" – refers to the subjective and pre-reflexive context in which a living being is immersed (Brentari, 2015: 75). In his works, Uexküll insists on the necessity of considering the organism as an entity inhabiting its own subjective environment, i.e. a world consisting of perceptual and operative signals endowed with pragmatic meanings (von Uexküll, 1982: 26-27;1928: 119; 1934: 45-46). Thus, the relationship between a living being and its surroundings is not simply mechanical, with the organism merely reacting to environmental stimuli through combinations of reflex arcs. Instead, each living being is at the centre of its own subjective environment (*Umwelt*), which is the result of its own perceptual capacities and interpretations of external and internal stimuli. Far from being reduced to a cartesian automaton, the organism constitutes an environment of pragmatic meanings (von Uexküll, 1982: 26-27).

Uexküll argues that the stimuli that the animal perceives through its species-specific sensory systems become signs of external objects that are objectified in the *Umwelt*. This semiotic activity of the organism thus consists in the synthesis and outwards transposition of stimuli that become signs of the presence of an object: sensory stimuli are transformed into neural patterns that produce signs of the external world (von Uexküll, 1928: 136-137). Brentari (2015: 111) describes this as a "transcendental biosemiotics", because the signs projected to the outside world do not have a denotative, but a constitutive and interpretative function. In conclusion, the external object is the synthetic unity of spatial, temporal, and qualitative signs that are the result of interpretation by the organism, which, even when not aware of this process, it is nevertheless the author of it.

As a contrast, the notion of *Umgebung* denotes the physical environment of the animal that, given its perceptual apparatus, does not enter in its *Umwelt*. It roughly corresponds, therefore, to the chemical and physical *milieu* that is not part of the organism's perceptual environment – although it can have a chemical and physical impact on it (Sharov & Tønnessen, 2021: 200).

It would be incorrect, however, to consider the notion of *Umwelt* only from the sensory side: Uexküll (1934: 39) clearly states that the subjective environment of an organism is the sum of its perceptual world (*Merkwelt*) and its operative world (*Wirkwelt*). The semiotic relationship of organism and environment, in fact, passes through the sense organs as well as the effectors (Farina & James, 2021: 423).

To illustrate this, Uexküll uses the theoretical model of the functional circuit, which describes how organisms associate certain perceptions with certain actions. We have already mentioned that the organism interprets sensory stimuli as signs of the presence of an object; in some cases, certain signs (*Merkzeichen*) work as perceptual marks (*Merkmale*) of an action-relevant object (von Uexküll, 1934: 47). At this point, the subject projects operative marks (*Wirkmale*) onto the object, i.e. action possibilities that set the effector organs in motion, which produce new effects on the world, in turn influencing the perceptual world and so on, following a circular model (von Uexküll, 1928: 119).

Uexküll adds that, in the case of higher animals, the meaning of a perceptual content is also determined by a certain "tone" – be it the emotional tone of the subject (*Stimmung*) or of the stimulus itself (*Ton*) (von Uexküll, 1982: 27-28; 1934: 104-105). This tone can determine the pragmatic meaning that the organism projects onto the object – its "operative image" (*Wirkbild*). The involvement of operative images in addition to perceptual and operative marks in the construction of the *Umwelt* implies the possibility for intraspecific semiotic variability – at least for animals capable of consciously perceiving objects in their environment.

In conclusion, with the concept of *Umwelt*, Uexküll emphasises the relational nature of the biological environment: there are as many *Umwelten* as there are living subjects (von Uexküll, 1928: 75). Thus, the environment is always the environment of a subject, and cannot be reduced to collection of neutral objects (von Uexküll, 1982: 27-28).

2. Two Interpretations of *Umwelt*

In this section, I will present the two main current interpretations of the concept of *Umwelt* that can be found in the cognitive sciences literature. To do so, I will rely on the analysis performed by Tim Feiten (2020). Each of the two variants will be useful in its own way both in explaining certain phenomena and in developing management strategies. For this reason, I will not favour one of the two notions to the detriment of the other, but I will retain this distinction.

2.1. *Umwelt* as a Selection

The fecundity of Uexküll's thought is now being rediscovered, especially by the advocates of the so-called "embodied cognition", a research programme that tries to uncover the bodily aspects of cognitive processes (Shapiro, 2007). The importance that he attributes to the bodily dimension in structuring the subjective environment is deemed relevant, because it shows the links between the organism's sensorimotor capacities and the affordances it can detect in its surroundings (Feiten, 2020). The concept of affordance is a relational one, since the same object can elicit different actions depending on the characteristics of its user (Chemero, 2009: 108). Thus, these authors consider Uexküll's ideas on the relationship between the *Bauplan* (bodily structure and functioning) and the *Umwelt* to be useful in emphasising the species-specific dimension of affordances (Baggs & Chemero, 2021: 2175).

According to these authors, the *Umwelt* denotes the subset of physical properties to which the organism has perceptual and operative access, by virtue of its species-specific sensorimotor endowment. For example, Baggs and Chemero (2021: 2178) state that "the physical realm is inherently meaningless, but the environment [the *Umwelt*] is not: the environment contains affordances". Further, they argue that the *Umwelt* is not a construction of the subject. Instead, it is a subset of the physical world carved out by the organism's sensorimotor capacities. The Uexküllian dynamic of meaning construction by the organism is missing here: the structure of the *Umwelt* is fixed and does not depend on the activity of the subject.

Dennett (2015:4-5) – an author not directly related to the embodied cognition perspective – also interprets the *Umwelt* as the portion of the physical world to which the organism has access via its sensory organs. According to his view, the active role played by the organism in the construction of meanings is missing: since there is no role for subjectivity in the construction of the *Umwelt*, the latter is reduced to the set of objects that a living being can perceptually discriminate.

In conclusion, whether it is to maintain the realism of the affordances of the ecological psychology tradition (Baggs and Chemero), or to avoid including the element of consciousness in animal cognition (Dennett), the interpretations presented here refer to the *Umwelt* as the product of a selection, by the sense organs, of an already given world. I call this "selectionist" *Umwelt*. Since it depends solely on its body structure, all members of a same species share the same selectionist *Umwelt*.

2.2. *Umwelt* as a Construction

Other interpreters (Campbell, Olteanu, Kull, 2019: 357; De Jesus, 2016; Feiten, 2020) place more emphasis on the productive aspect of the notion of

Umwelt: it is the organism that constructs its world of pragmatic meanings through the establishment of functional circuits. In this way, these authors can identify the dynamics of the constitution of affordances, which, although perceived by the animal as objective properties of the world, are in fact the result of semiotic construction (Feiten, 2020: 7).

In section §1 we saw how, for Uexküll, affordances are something constructed by the organism: first, perceptual marks are linked to operative marks according to the functional circuit model. Moreover, the concepts of stimulus tone (*Ton*) and emotional tone (*Stimmung*) express both the idea that perception is always characterised by certain calls to action and the thesis that the operative image attributed to a perceptual content depends on the emotional state of the animal. I call this "constructionist" *Umwelt*[1]. Unlike the selectionist one, this variant can account for intraspecific differences between *Umwelten*.

The distinction between the selectionist and constructionist notions of *Umwelt* should not, however, lead us to think that the two interpretations are incompatible. On the contrary, they should be understood as complementary. In fact, in section §1 we saw both how an organism's perceptual and operative world depends systematically on its anatomical and physiological structures, and how the activity of functional circuits and the *Stimmung* leads the subject to apply pragmatic meanings to surrounding objects. In addition, Uexküll himself seems to oscillate between a selectionist and a constructionist sense of *Umwelt* (Feiten, 2020: 3-4).

The reason why I keep this distinction is twofold: first, I want to investigate how both uses may be useful in explaining certain (adaptive and non-adaptive) behaviours exhibited by animals in contact with urbanisation phenomena (§3); second, I want to elaborate some guidelines for possible management strategies for urban fauna (§4). In both these attempts, the maintenance of the distinction between selectionist and constructionist *Umwelt* allows us for a more fine-grained analysis of the phenomena examined, as well as a better understanding to which senses of subjective environment are employed in the literature I will refer to.

[1] This view can be also called "organismic-centered", because it challenges the traditional perspective of an objective environment valid for all living organisms (Farina & Belgrano, 2006: 7). In this regard, the concept of eco-field has been proposed (Farina & Belgrano, 2006), understood as a "spatial configuration with a specific meaning-carrier for every semethic organism-resource interaction" (Sánchez-García et al., 2017: 58).

3. *Umwelt* and Urban Ecology: Explanatory Usefulness

Ever since the development of agriculture, human beings have made significant modifications to their natural surroundings: the use of soil, the creation or modification of watercourses, the cultivation of certain plants to the detriment of others are just a few examples of our ecosystem engineering (Casetta, 2023: 75; Chu & Karr, 2013). Such alterations have often had strong consequences on biodiversity, leading to the reduction in fitness of some species and the proliferation of others (Boivin et al., 2016; Johnson & Munshi-South, 2017).

Of all anthropogenic interventions, urbanisation is undoubtedly among those that modify pre-existing ecosystems in the most radical ways, altering the availability of resources and generating new heterogeneous spaces (roads, parks, buildings, poles), in which natural elements are often mixed with artificial ones (Farina, 2020: 22-23). This makes cities a major source of biodiversity disturbance (Guetté et al., 2017: 139), creating new ecological opportunities for some species (Griffin, Netto, Peneaux, 2017: 15; Lowry, Lill, Wong, 2012: 538; Toger et al., 2018) and dangers for others (Egerer & Buchholz, 2021: 2255; Robertson, Rehage, Sih, 2013).

Over the last few decades, there has been a rapid increase both in the number of cities and in their size (Dunn et al., 2022: 1), followed by the incorporation into urban contexts of species that were previously unaccustomed to cities (Niesner et al., 2021: 3; Toger et al., 2018). For these reasons, there is a need to investigate how living species relate not only to the physical and chemical changes in cities (Casetta, 2023: 111), but also to the radical changes in stimuli and affordances that are inevitably brought about by urbanisation.

Since the analyses elaborated in this section are inevitably affected by a large degree of generality, a clarification is needed. In this section I confine myself to discussing the way in which the two senses of the concept of *Umwelt* can help us interpret certain phenomena, occasionally giving a few examples. Clearly, a full study of the phenomena relating to urban biodiversity would require a complementary in-depth examination of at least two fundamental factors: (1) the peculiar characteristics of the urban biodiversity present (such as species, quantity of specimens, population density, inter- and intraspecific relationships, degree of intraspecific variability[2], etc.), (2) the material (physical-chemical, geographical) and

[2] For reasons that will become clear in §3.2, cities can be privileged contexts for measuring intraspecific variability (Harding et al., 2019).

cultural characteristics (citizens' behaviour, cultural perception of the species etc.) of the urban context examined.

3.1. Explanatory usefulness of selectionist *Umwelt*

Since it involves significant alterations of the pre-existing nature, the phenomenon of urbanisation is recognised as one of the most relevant high-speed changes to the natural environment conducted by human beings (McKinney, 2002; Shochat, Warren, Faeth, 2006; Sol, Lapiedra, González-Lagos, 2013: 1101): consequently, cities introduce sets of new risks and resources for which the pre-existing species on the territory have not evolved accordingly. Such "urban ecological novelties" (Zuñiga-Palacios et al., 2021: 2) can therefore favour the success of some species and the disappearance of others from the given geographical area, as well as the incorporation of allochthonous species – either because they are attracted by the resources of the new urban context, or as a consequence of the increase in human movements due to globalisation – which, in turn, constitutes an ecological novelty for native species (Robertson, Rehage, Sih, 2013). In this scenario, the species-specific perceptual capacities, that enable the organism to detect environmental signals useful for survival, assume fundamental importance in determining its adaptive success or failure. In this regard, the selectionist *Umwelt* can play an important analytical role for two main reasons.

First, the outlining of the selectionist *Umwelt* of a given species can point to the perceptual blindness underlying certain maladaptive behaviours in urban contexts. For example, it is estimated that, each year, millions of birds die due to collisions with the windows of urban buildings (Farina, 2020: 27). An important role in explaining this phenomenon is played by the inability of many avian species to perceive bright, reflective surfaces (Klem, 1989: 616). In this case, the reflective property of windows is not present in the birds' *Umwelten*, but rather is confined to their *Umgebungen*. On the other hand, the *Umgebung* – understood as the physical-chemical environment to which the animal does not perceive – can exert detrimental effects on the quality of cognitive processes, as in the case of pesticides causing spatial working memory impairment in bees (Samuelson et al., 2016).

Second, the reconstruction of a species' selectionist *Umwelt* can serve not only to determine which objects and/or physical-chemical stimuli are not part of its perceptual world, but also to identify those behavioural performances caused by the perception of species-specific affordances. This concern, for example, cases in which certain sections of buildings are operatively perceived as nesting or resting places because of their shape and position (Niesner et al., 2021: 3; Zuñiga-Palacios et al., 2021: 7), or when birds begin to sing before dawn due to artificial lighting, sometimes increasing their reproductive success, as in the case of male Blue tits (Sol, Lapiedra, González-Lagos, 2013: 1107).

Sometimes, however, environmental signals that work as affordances at a species-specific level can act as ecological traps, i.e. where a species makes use of resources that reduce its reproductive success (fitness) (Robertson & Hutto, 2006; Hale & Swearer, 2016). This is because the attribution of operative meanings to stimuli is often genetically determined (Farina, 2020: 31), i.e. the result of the evolutionary pathway of the species in question. In cities, however, it happens that species-specific operative responses turn out to be counterproductive, because such behavioural traits are activated in environmental conditions that are very different from those that had favoured their evolutionary selection (Hale, Morrongiello, Swearer, 2016: 1).

In Uexküllian terms, this happens due to two types of phenomena: when a signal attracts a species to a city where the unperceived physical-chemical conditions (*Umgebung*) are detrimental to that population, or when the organism misinterprets a stimulus as a sign of the presence of an object in its natural environment.

An example of the first type concerns certain species of fish that, attracted by the milder water temperatures, come to urban canals in winter, but encounter a reduction in numbers and reproductive capacity due to certain pollutants (Zuñiga-Palacios et al, 2021: 6). Another example concerns bats that, attracted by the cavities of urban buildings and the climate of cities, end up being electrocuted by contact with high-voltage cables (Zuñiga-Palacios et al. 2021: 7) or preyed upon by domestic cats (Ancillotto, Serangeli, Russo, 2013).

Cases belonging to the second type are, on the other hand, Cuban tree frogs (*Osteopilus septentrionalis*) ingesting Christmas lights bulbs, because they are similar to the bioluminescent activities of natural prey (Robertson, Rehage, Sih, 2013: 553), or the laying of eggs by dragonflies on glass or solar panels that horizontally polarise light, a perceptual signal that these organisms use to locate water mirrors (Robertson & Blumstein, 2019: 5).

3.2. Explanatory Usefulness of Constructionist *Umwelt*

I have shown that, in urbanisation phenomena, the numerous alterations of the material environment correspond to important changes in the landscape of signals that are informative for the behaviour of urbanised living beings. If, on the one hand, most operative responses to urban affordances are fixed at the species-specific level – in other words, at the genetic level (Farina, 2020: 31) – on the other hand, some populations of a certain species may adapt to new urban contexts by developing new operative responses to certain signals, or by learning to attribute meanings to stimuli that, in the natural context in which their species evolved, are absent or of no practical interest. Such creation of new meanings (or reconfiguration of those already available) can lead to the adaptive success of a species in the urban context, sometimes with an even higher rate of reproduction than conspecifics present in the original ecological niches: this phenomenon falls under the name of "synurbization" (Francis & Chadwick, 2012).

Given its emphasis on the subjective construction of meanings by the individual organism, the constructionist notion of *Umwelt* may represent a useful explanatory tool for examining the semiotic nature of such phenotypic plasticity (Casetta, 2023: 69). Indeed, we can distinguish two general phenomena underlying the creation of new pragmatic meanings: the alteration of *Stimmung* and learning. As it will become clear, these two factors are often interrelated.

The urban context can significantly affect the emotional condition of the living beings that inhabit it, whether these are humans (De Franco & Moroni, 2023: 4) or animals (Lowry, Lill, Wong, 2012: 539). This can consist of either the effect of chemical pollutants that can alter mood at the hormonal level (Wojnarowski et al, 2021: 8), or in the presence/absence of certain stimuli that the subject is able to perceive (such as noise pollution) (Halfwerk & Slabbekoorn, 2015: 5) and that are sometimes loaded with pragmatic meaning (such as the presence of food, or the absence of predators) (Łopucki, Klich, Kiersztyn, 2021: 8). The most widespread *Stimmung* alterations consist in stress levels increase, greater boldness towards surrounding risks, and higher degrees of aggression or sociability compared to non-urbanised conspecifics.

In these cases, the organism's emotional tone leads to the attribution of new operative images (*Wirkbilder*) to certain surrounding elements (objects and other animals, including human individuals) or to "colour" the environment with a general operative tone (Sharov & Tønnessen, 2021: 139). Examples of the first case are increased aggression towards conspecifics during high levels of stress (Kekkonen, 2017: 229), or tolerance behaviour towards other individuals near food sources, when the abundance of resources ensures a relatively constant sense of satiety (Łopucki, Klich, Kiersztyn, 2021: 2). Examples of the second case are alterations in circadian rhythms (Halfwerk & Slabbekoorn, 2015: 5) and anticipation of reproductive timing

(Lowry, Lill, Wong, 2012: 540) due to stress, or increased exploratory behaviour of the surroundings due to boldness[3] (Thompson et al., 2018: 1415).

It is important to note that the increased bold behaviour of city fauna – one of the most frequently detected phenomena in the urban ecology literature (Lowry, Lill, Wong, 2012: 539) – can be significantly linked to the second mode of meaning-making I mentioned earlier, namely learning. On the one hand, in fact, the increased exploration of new areas by the bolder specimens is also an opportunity for these individuals to gather more information (Thompson et al., 2018: 1422) and, consequently, to elaborate new semiotic relationships between the environmental elements at hand. On the other hand, boldness itself may be the result of familiarisation with certain urban elements (Uchida et al., 2019: 1584), such as the replacement of the operative image of escape with that of approachability towards humans. Behaviours due to increased tolerance and sociability may also favour the transmission of new meanings between conspecifics (Dimitras, Ross, Stegman, 2021: 17; Greggor et al., 2014: 493), as well as between individuals of different species (Lefebvre & Boogert, 2010: 126).

In the field of biosemiotics, learning is understood as the acquisition and/or modification of new relationships between signs and is often referred to as "semiogenesis" (Sharov & Tønnessen, 2021: 248). In this process, the network of (perceptual and operative) meanings available to the organism takes on a new configuration (Campbell, Olteanu, Kull, 2019: 356; Kull, 2018: 139). Uexküll himself already argues for the need to consider phenomena of semiogenesis (von Uexküll, 1928: 9) – although he does not devote much space to them[4] – as in many animal species there is a semiotic variability that is proportionate to the complexity of their bodily structures called *Baupläne* (Brentari, 2015: 142).

The nature of the new relationships between signs depends on the type of learning mechanism brought into play: for example, learning by conditioning (pavlovian or operant) consists, at the semiotic level, in constructing a sign of spatio-temporal contiguity between two events called an "index". Imitative learning consists in using the identity sign to link the

[3] In behavioural terms, boldness can be defined as "the manner in which an individual/population respond to threatening situations: the boldest individuals being willing to take more risks" (Lowry, Lill, Wong, 2012: 3).

[4] The reason why Uexküll focuses little on the acquisition of new relationships between signs is to be found in his thesis of the harmony and fixity of the relationship between the organism and its environment. Uexküll's thought lacks a systematic reflection on cases in which alterations in the *Umgebung* of a species lead to a change in its *Umwelt* (Tønnessen, 2009).

perceptual signs of conspecifics' actions with operant signs (Kull, 2018: 140-141).

In the case of urban species, learning enables them to cope with those ecological novelties that do not fit into the perception-action patterns already arranged at the level of their selectionist *Umwelt*[5]. One of the most widespread forms of urban species' learning is the familiarisation with certain environmental elements previously recognised as dangerous (Uchida et al., 2019: 1584). Examples of this phenomenon are countless, from interactions with humans to receive food from squirrels (Uchida et al., 2019: 1588) or sparrows (Dimitras, Ross, Stegman, 2021: 31), to reducing the minimum safe distance used by some squirrel species before activating escape behaviour (Uchida et al., 2019: 1588). Moreover, familiarisation can sometimes be transferred from one set of signals to another, as in the case of urban Eastern grey squirrels (*Sciurus carolnensis*), that become less frightened by domesticated dogs because they are accompanied by humans (who often keep them on a leash and provide them with food) (Uchida, 2019: 1587-1588).

Other types of phenomena due to modification or creation of new meanings include: the ability to adapt one's behaviour to specific human individuals (Levey et al., 2009; Sol, Lapiedra, González-Lagos, 2013: 1108); the modification of activity patterns based on the rhythms of the city and the fragmentation of its spaces (Niesner et al, 2021: 4)[6]; the ability to establish new operative images with artificial objects, as in the case of Blue tits that learn to open milk bottles (Lefebvre & Boogert, 2010).

Considering these reconfigurations of meanings allows us to explain why urban specimens may show considerable behavioural variation from their conspecifics while genetic adaptations are relatively rare (Lowry, Lill, Wong, 2012: 539). Furthermore, the constructionist notion of *Umwelt* allows us to understand that, at some times, it is variations in operative meanings within a population that make long-term evolutionary changes possible at the phylogenetic level. The construction of pragmatic meanings leads to changes in the behaviour and relationships the organism has with its surroundings, which in turn can retroact on the organism by imposing a further modification of its *Umwelt* or genetic make-up (through natural selection). Such a form of cyclical causation between the organism and its environment shows how organisms can play an active role in determining their evolutionary pathway (Casetta, 2023: 76), and is a useful explanatory model for understanding the

---

[5] For a detailed list of animal learning mechanisms see Greggor et al. (2014: 491-493); for an analysis of the semiotic aspects of such phenomena see Kull (2018: 139-141).

[6] See Farina & Villa (2023) for a model of animal semiotic coding of human-made sounds.

transition from behavioural changes concerning ontogenetic development (Omenn & Motulsky, 2006: 15-16) – the reversibility of which was detected during the Covid-19 lockdown (Gordo et al, 2021) – to those that are irreversible because they are genetically fixed and transmitted to offspring (Perrier, Caizergues, Charmantier, 2020; Kull, 2018: 290) .

The mechanism of natural selection is not always the best explanation for many adaptive behaviours that urban populations perform in response to important ecological novelties in urban contexts: sometimes, in fact, behavioural changes occur due to semiogenesis phenomena, which are followed by changes at the level of the ecological niche, which eventually exerts selective pressures at the genetic level (Kull, 2018: 287)[7]. Although it is difficult to empirically demonstrate the intervention of semiotic factors within the evolutionary dynamics of a species, such a hypothesis can be useful in explaining both the presence of effective adaptations not genetically based, and the way in which the prolonged presence of a species in urban contexts can lead to changes on a phylogenetic scale. An example of this is how the widespread use by urban House finches (*Carpodacus mexicanus*) of the food left to them by humans has resulted in selective pressures that ultimately favoured changes in the morphology of their beaks (Sol, Lapiedra, Ducatez, 2020: 258).

It is therefore no coincidence that Uexküll's thinking has been taken up by the proponents of the so-called "Extended Evolutionary Synthesis", according to which evolution consists not only in the genetic adaptation of a species to its environment by means of natural selection, but also in the active construction by organisms of their ecological niches and phenotypes (Laland, Matthews, Feldman, 2016; Casetta, 2023: 60-61).

4. *Umwelt* and Urban Ecology: Usefulness for Urban Species Management

In recent years, the philosophical debate in ecology has focused on some main concepts: "Anthropocene", according to which we live in a new geological era caused by the human bio-geo-physical impact on the environment (Crutzen, 2006); "Gaia hypothesis", according to which the Earth is a superorganism that, through the activity of the biosphere, tends homeostatically towards dynamic equilibrium (Lovelock & Margulis, 1974);

[7] This semiotic mechanism of evolution can be equated with the so-called "Baldwin effect", in which epigenetic changes precede genetic changes (Kull, 2018; Sharov & Tønnessen, 2021: 236).

"planetary boundaries", that denotes the thresholds of impact on the planet beyond which the human activity on Earth would be unsustainable (Tønnessen, 2020: 94). It should be noted that all three notions are located at a physical-chemical level and do not refer to the dynamics of meaning between living beings and their environment: this is arguably a symptom of an ecology that usually describes ecosystem dynamics only as exchanges of matter and energy made possible by trophic chains (Farina, 2014: 3; Tønnessen, 2020: 94).

The necessary complement to this approach is the development of an "ecosemiotic" perspective (Farina & James, 2021; Tønnessen, 2020) able to address the semiotic aspects of ecological dynamics by considering the *Umwelten* and their dynamics of construction and/or modification[8]. In the previous section, I have shown how the semiotic perspective inaugurated by Uexküll contains useful analytical resources for shedding light on the informational nature of urban contexts. In this section I will show how the two interpretations of *Umwelt* can be valuable not only as explanatory tools, but also as basis for developing urban fauna management strategies.

Such an approach can sometimes be a valuable alternative to the most widespread interventions operating at the chemical level, such as disinfestation with poison (Ferretti & Chiaranz, 2021: 10). This kind of intervention is, in fact, limited in several aspects. First, it can be dangerous for non-targeted species and/or for the environment: for instance, there are numerous cases of cats ingesting mouse baits or eating poisoned mice (Ferretti & Chiaranz, 2021: 16), as there are numerous pathologies (neurodegenerative, of the endocrine systems, tumoral) affecting many species (humans included), that are favoured by the excessive use of chemical anti-mosquito agents (Agnelli et al., 2015: 11-12). Second, such solutions are usually only partial and short-term, since they often fail to prevent the return and proliferation of the infesting species (Ferretti & Chiaranz, 2021: 10). Finally, in many cases chemical disinfestation can cause great suffering to the affected organism, as in the case of rodenticides that operate through anti-coagulation (Ferretti & Chiaranz, 2021: 16).

In conclusion, if we want preventive and long-term solutions, we should rethink the strategies for managing urban fauna by manipulating, where

[8] In this paper I focus exclusively on the alterations of the non-human animals' *Umwelten*. It should be noted, however, that a sociosemiotic investigation on human inhabitants's perception of the other urban species is a necessary complement for the description of local semiotic dynamics. See Maran (2014) for the semiotic concept of "locality", and Magnus, Remm, Kull (2024) for a case study in four Estonian towns.

possible, the meanings that the city offers to its inhabitants, and relegate chemical solutions to a corollary of a semiotic approach.

4.1. Selectionist *Umwelt* and Urban Species Management

In section 3.1, I showed how the specification of the selectionist *Umwelt* of a species allows us to identify not only the operative meanings offered by the urban context, but also those maladaptive behaviours that are due to perceptual blindness. The literature already offers some proposals for interventions at the semiotic level, and some of these are already used in urban fauna management.

A first example is the use of glass barriers on the edges of motorways to reduce noise pollution – a factor that can cause high levels of stress for some species – decorated with silhouettes of falcons to reduce the risk of collision by local avifauna (Farina, 2020: 28). In doing so, the risk of birds crashing into glass surfaces due to their perceptual limitations is corrected by inserting environmental signals that, for those species, have an operative image of escape. Similarly, it is possible to limit unwanted behaviours caused by species-specific affordances with other signals endowed with operative meaning, such as the use of silhouettes of falcons or structures that resemble the eye of predators, to fight pigeons nesting in building cavities (Ferretti & Chiaranz, 2021: 37).

There are also several proposals to reduce the cases of ecological traps due to the erroneous interpretation of certain perceptual signals. One of them regards the phenomenon of dragonfly oviposition on glass that is mistaken for water surfaces (§3.1) and consists in adding white bands to the surface (Robertson, Rehage, Sih, 2013: 557). Ecological traps can be counteracted by manipulating the target species' behaviour by using other affordances, or by physically limiting the access to the trap (Robertson & Blumstein, 2019: 5). An example of the latter option is the use of anti-bird nets on artificial carp breeding ponds, to prevent some species – such as the Red-necked grebe (*Podiceps grisegena*) – from coming there to reproduce. These birds, in fact, initially attracted by the availability of small carps, become unable to feed the offspring due to the excessive size of these fishes (grown up by the time of the eggs hatching) (Robertson & Blumstein, 2019: 4).

4.2. Constructionist *Umwelt* and Urban Species Management

The first and most important pragmatic implication of the constructionist notion of *Umwelt* consists in the preliminary recognition that in some species there can be a certain intraspecific variability in their operative meanings. Several types of precautions and possibilities for interventions follow from this consideration.

First, since intraspecific variation is made possible by the way in which the organism's *Stimmung* alters the operative images of perceptual contents, it becomes possible to influence the behaviours of an urban species by acting on those elements that cause changes in emotional condition. Of this kind are the interventions aimed at reducing stress from noise pollution, such as the glass barrier on the sides of highways that I mentioned before (§4.1).

Furthermore, since many phenomena of intraspecific variation are due to learning, the constructionist *Umwelt* allows us to explain why some species-specific semiotic solutions (§3.1) can lose efficacy – as in the case of birds that habituate to scarecrows (Marsh, Erickson, Salmon, 1992).

It also seems possible to identify those elements of the city that easily enter the *Umwelten* of urban animals through learning, such as building overhangs, waste collection areas, waterways and so on (Niesner et al., 2021: 5). However vague and hypothetical, the knowledge of the traits that facilitate the creation of operative meanings by urban species can be helpful for a more conscious management. Proof of this is the human introduction of the Peregrine falcon (*Falco peregrinus*) into cities. This species learned to exploit the features of the urban context to nest and obtain food: this led to the regrowth of the population after it had been severely reduced by the use of DDT (Mak, Francis, Chadwick, 2021)[9].

Finally, it is possible to exploit social learning dynamics for conservation goals, like promoting opportunities for certain individuals to meet conspecifics capable of performing a target behaviour. For example, one could try to reduce the number of birds' collisions with street lamps by initially training some individuals in a flock; subsequently, the reintroduction of the trained individuals into the flock could favour the diffusion of this knowledge to conspecifics (even to other flocks) (Greggor et al., 2014: 493-494), as well as to individuals of other species (Lefebvre & Boogert, 2010: 126).

---

[9] Moreover, the introduction of peregrine falcons in urban areas may be helpful in controlling pigeon populations.

Conclusion

In this article I have used a semiotic perspective to explain some types of adaptive and maladaptive behaviours of urban fauna. I showed how the distinction between the selectionist and the constructionist interpretations of *Umwelt* can be useful both as an explanatory concept and a pragmatic tool for management interventions. These considerations encourage the development of a kind of urban ecology capable of integrating the physical-chemical level of analysis with the sensorial and semiotic one (Farina, Krause, Mullet, 2024). Roads, buildings, waterways etc. are not only physical and quantitatively measurable places, but also contexts of experiences and meanings: this is valid for humans (De Franco & Moroni, 2023) as well as for the many living beings that live with us in the city.

Urban development is a rapidly increasing process: it is estimated that, by 2030, approximately 75% of humanity will live in cities. For these reasons, the loss of biodiversity, the urbanisation of wild and/or invasive species and the interactions between species (with the related risks of spreading pathogens) are phenomena destined to increase (Ferretti & Chiaranz, 2021: 149). This makes the development of an adequate urban ecology one of the most important challenges of our present. To do so, we need appropriate conceptual tools to implement effective management policies: the notion of *Umwelt* is certainly among the notions that must figure in this complex conceptual work.

References

Agnelli, P., Bellucci, V., Bianco, P. M., Campanella, L., Jacomini, C., Modonesi, C. M., Panizza, C., & Tamino, G., (2015), *Impatto sugli esseri viventi e sugli ecosistemi delle sostanze sintetiche utilizzate nella profilassi anti-zanzara*, ISPRA, Roma.

Ancillotto, L., Serangeli, M. T., & Russo, D., (2013), "Curiosity Killed the Bat: Domestic Cats as Bat Predators", in *Mammalian Biology*, pp. 369-373.

Baggs, E. & Chemero, A., (2021), "Radical Embodiment in Two Directions", in *Synthese*, 198 (9), pp. 2175-2190.

Boivin, N. L., Zeder, M. A., Fuller, D. Q., Crowther, A., Larson, G., Erlandson, J. M., Denham, T., & Petraglia, M. D., (2016), "Ecological Consequences of Human Niche Construction: Examining Long-Term

Antrhopogenic Shaping of Global Species Distribution", in *PNAS*, 113 (23), pp. 6388-6396.

Brentari, C., (2015), *Jakob Von Uexküll. The Discovery of the Umwelt between Biosemiotics and Theoretical Biology*, Springer, Dordrecht.

Campbell, C., Olteanu, A., & Kull, K., (2019), "Learning and Knowing as Semiosis: Extending the Conceptual Apparatus of Semiotics", in *Sign Systems Studies*, 47 (3), pp. 352-381.

Casetta, E., (2023), *Filosofia dell'ambiente*, Il Mulino, Bologna.

Chemero, A., (2009), *Radical Embodied Cognitive Science*, MIT Press, Cambridge (MA).

Chu, E. W. & Karr, J. R., (2013), "Environmental Impact, Concept and Measurement of", in *Encyclopedia of Biodiversity*, 3, pp. 278-296.

Crutzen, P. J., (2006), "The 'Anthropocene'", in Ehelers E. & Krafft T. (eds.), *Earth System Science in the Anthropocene*, Springer, Berlin, pp. 13-18.

De Franco, A. & Moroni, S., (2023), "The City as an Information System: Urban Agency, Experiential Inputs and Planning Measures", in *Cities*, 134 (104183), pp. 1-8.

De Jesus, P., (2016), "From Enactive Phenomenology to Biosemiotic Enactivism", in *Adaptive Behavior*, 24 (2), pp. 130-146.

Dennett, D. C., (2015), "Why and How Does Consciousness Seem the Way it Seems?", in Metzinger T. & Windt J. M. (eds.), *Open MIND*, MIND Group, Frankfurt.

Dimitras, E., Ross, A., & Stegman S., (2021), *Investigating the Welfare of Wild Animals in Urban Environments*, Animal Ethics, Oakland.

Dunn, R. R., Burger, J. R., Carlen, E. J., Koltz, A. M., Light, J. E., Martin, R. A., Munshi-South, J., Nichols, L. M., Vargo, E. L., Yibarek, S., Zhao, Y., & Cibrián-Jaramillo, A., (2022), "A Theory of City Biogeography and the Origin of Urban Species", in *Frontiers in Conservation Science*, 3 (761449), pp. 1-18.

Egerer, M. & Buchholz, S., (2021), "Reframing Urban 'Wildlife' to Promote Inclusive Conservation Science and Practice", in *Biodiversity and Conservation*, 30, pp. 2255-2266.

Farina, A., (2014), "From Umwelt to Soundtope: an Epistemological Essay on Cognitive Ecology", in *Biosemiotics*, 7, pp. 1-10.

Farina, A., (2020), "Hybrid Nature: Effects on Environmental Fundamentals and Species' Semiosis", in *Biosemiotics*, 13, pp. 21-40.

Farina, A. & Belgrano, A., (2006), "The Eco-Field Hypothesis: Toward a Cognitive Landscape", in *Landscape Ecology*, 21, pp. 5-17.

Farina, A. & James, P., (2021), "Vivoscapes: an Ecosemiotic Contribution to the Ecological Theory", in *Biosemiotics*, 14 (2), pp. 419-431.

Farina, A., Krause, B., & Mullet, T. C., (2024), "An Exploration of Ecoacustics and its applications in Conservation Ecology", in *BioSystems*, 245 (105296).
Farina, A. & Villa, A. E., (2023), "On the Semantics of Ecoacoustic Codes", in *Byosystems*, 232 (105002).
Feiten, T. E., (2020), "Mind after Uexküll: a Foray into the World of Ecological Psychologists and Enactivists", in *Frontiers in Psychology*, 11 (480), pp. 1-10.
Ferretti, S. & Chiaranz, G., (2021), *Manuale di gestione della fauna urbana. Approccio alla biodiversità e all'ecologia in città*, Il Piviere Edizioni, Gavi.
Francis, R. A. & Chadwick, M. A., (2012), "What Makes a Species Synurbic?", in *Applied Geography*, 32, pp. 514-521.
Gordo, O., Brotons, L., Herrando, S., & Gargallo, G., (2021), "Rapid Behavioural Response of Urban Birds to COVID-19 Lockdown", in *Proceedings in Biological Sciences*, 288 (20202513), pp. 1-8.
Greggor, A. L., Clayton, N. S., Phalan, B., & Thornton, A., (2014), "Comparative Cognition for Conservationists", in *Trends in Ecology and Evolution*, 29 (9), pp. 489-495.
Griffin, A. S., Netto, K., & Peneaux, C. (2017), "Neophilia, Innovation and Learning in an Urbanized World: a Critical Evaluation of Mixed Findings", in *Current Opinion in Behavioral Sciences*, 16, pp. 15-22.
Guetté, A., Gaüzère P., Devictor, P., Jiguet, F., & Godet L., (2017), "Measuring the Synanthropy of Species and Communities to Monitor the Effects of Urbanization on Biodiversity", in *Ecological Indicators*, 79, pp. 139-154.
Hale, R., Morrongiello, J. R., & Swearer, S. E., (2016), "Evolutionary Traps and Range Shifts in a Rapidly Changing World", in *Biology Letters*, 12 (6), pp. 1-4.
Hale, R. & Swearer, S. E., (2016), "Ecological Traps: Current Evidence and Future Directions", in Proceedings of the Royal Society B: Biological Sciences, 283 (20152647).
Halfwerk, W. & Slabbekoom, H., (2015), "Pollution Going Multimodal: the Complex Impact of the Human-Altered Sensory Environment on Animal Perception and Performance", in *Biology Letters*, 11 (20141051), pp. 1-7.
Harding, H. R., Gordon, T. A. C., Eastcott, E., Simpson, S. D., & Radford, A. N., (2019), "Causes and Consequences of Intraspecific Variation in Animal Responses to Anthropogenic Noise", *Behavioral Ecology*, 30 (6), pp. 1501-1511.

Johnson, M. T. J. & Munshi-South, J., (2017), "Evolution of Life in Urban Environments", in *Science*, 358 (607).

Kekkonen, J., (2017), "Pollutants in Urbanized Areas: Direct and Indirect Effects on Bird Populations", in Murgui E. & Hedblom M. (eds.), *Ecology and Conservation of Birds in Urban Environments*, Springer, Cham.

Klem, D., (1989), "Bird: Window Collisions", in *The Wilson Bulletin*, 101 (4), pp. 606-620.

Kull, K., (2018), "On the Logic of Animal Umwelten: the Animal Subjective Present and Zoosemiotics of Choice and Learning", in Marrone G. & Mangano D. (eds.), *Semiotics of Animals in Culture*, Springer, Cham, pp. 135-148.

Laland, K., Matthews, B., & Feldman, M. W., (2016), "An Introduction to Niche Construction Theory", in *Evolutionary Ecology*, 30, pp. 191-202.

Lefebvre, L. & Boogert, N. J., (2010), "Avian Social Learning", in *Encyclopedia of Animal Behavior*, 1, pp. 124-130.

Levey, D. J., Londoño, G. A., Ungvari-Martin, J., Hiersoux, M. R., Jankowski, J. E., Poulsen, J. R., Stracey, C. M., & Robinson, S. K., (2009), "Urban Mockingbirds quickly Learn to Identify Individual Humans", in *PNAS*, 106 (22), pp. 8959-8962.

Łopucki, R., Klich, D., & Kiersztyn, A., (2021), "Changes in the Social Behavior of Urban Animals: more Aggression or Tolerance?", in *Mammalian Biology*, 101, pp. 1-10.

Lovelock, J. E. & Margulis, L., (1974), "Atmospheric Homeostasis by and for the Biosphere: the Gaia Hypothesis", in *Tellus*, 26 (1-2), pp. 2-10.

Lowry, H., Lill, A., & Wong, B. B. M., (2012), "Behavioural Responses of Wildlife to Urban Environments", in *Biological Reviews*, 88, pp. 537-549.

Magnus, R., Remm, T., & Kull K., (2024), "Semiotic Space for Native Biota in the City", in Bellentani F., Panico M., Yoka L. (eds.), *Semiotic Approaches to Urban Space. Signs and Cities*, Edwar Elgar Publishing, Cheltenham, pp. 193-208.

Mak, B., Francis, R. A., & Chadwick, M. A., (2021), "Breeding Habitat Selection of Urban Peregrine Falcons (Falco Peregrinus) in London", in *Journal of Urban Ecology*, 7 (1), pp. 1-9.

Maran, T., (2014), "Place and Sign. Locality as a Foundational Concept for Ecosemiotics", in Siewers A. K. (ed.) *Re-Imagining Nature. Environmental Humanities and Ecosemiotics*, Bucknell University Press, Plymouth, pp. 79-89.

Marsh, R. E., Erickson, W. A., & Salmon, T. P., (1992), "Scarecrows and Predator Models for Frightening Birds from Specific Areas", in *Proceedings of the Fifteenth Vertebrate Pest Conference*, 15, pp. 112-114.

McKinney, M. L., (2002), "Urbanization, Biodiversity and Conservation", in *BioScience*, 52, pp. 883-890.

Niesner, C. A., Blakey, R. V., Blumstein, D. T., & Abelson, E. S., (2021), "Wildlife Affordances of Urban Infrastructure: a Framework to Understand Human-Wildlife Space Use", in *Frontiers in Conservation Science*, 2 (7774137), pp. 1-7.

Omenn, G. S. & Motulsky, A. G., (2006), "Ecogenetics: Historical Perspectives", in Costa, L. G. & Eaton, D. L. (eds.), *Gene-Environment Interactions: Fundamentals of Ecogenetics*, John Wiley and Sons, Hoboken (NJ), pp. 7-16.

Perrier, C., Caizergues, A., & Charmantier, A., (2020), "Adaptation Genomics in Urban Environment", in Szulkin M., Mushi-South J., & Charmantier A. (eds.), *Urban Evolutionary Biology*, Oxford University Press, Oxford.

Robertson, B. A. & Blumstein, D. T., (2019), "How to Disarm an Evolutionary Trap", in *Conservation Science and Practice*, 1 (11), pp. 1-14.

Robertson, B. A. & Hutto, R. L., (2006), "A Framework for Understanding Ecological Traps and an Evaluation of Existing Evidence", in *Ecology*, 87 (5), pp. 1075-1085.

Robertson, B. A., Rehage, J. S., & Sih, A., (2013), "Ecological Novelty and the Emergence of Evolutionary Traps", in *Trends in Ecology and Evolution*, 28 (9), pp. 552-560.

Samuelson, E. E. W., Chen-Wishart, Z. P., Gill, R. J., & Leadbeater, E., (2016), "Effect of Acute Pesticide Exposure on Bee Spatial Working Memory Using an Analogue of the Radial-Arm Maze", in *Scientific Reports*, 6 (38957), pp. 1-11.

Sánchez-García, F. J., Machado, V., Galián, J., & Gallego, D., (2017), "Application of the Eco-Field and General Theory of Resources to Bark Beetles: Beyond the Niche Construction Theory", in *Biosemiotics*, 10, pp. 57-73.

Shapiro, L., (2007), "The Embodied Cognition Research Programme", in *Philosophy Compass*, 2 (2), pp. 338-346.

Sharov, A. & Tønnessen, M., (2021), *Semiotic Agency. Science beyond Mechanism*, Springer, Cham.

Shochat, E., Warren, P. S., & Faeth, S. H. H., (2006), "Future Directions in Urban Ecology", in *Trends in Ecology and Evolution*, 21, pp. 661-662.

Sol, D., Lapiedra, O. & Ducatez, S., (2020), "Cognition and Adaptation to Urban Environments", in Szulkin M., Mushi-South J., Charmantier A. (eds.), *Urban Evolutionary Biology*, Oxford University Press, Oxford.

Sol, D., Lapiedra, O., & González-Lagos, C., (2013), "Behavioural Adjustments for a Life in the City", in *Animal Behaviour*, 85, pp. 1101-1112.

Thompson, M. J., Evans, J. C., Parsons, S., & Morand-Ferron, J., (2018), "Urbanization and Individual Differences in Exploration and Plasticity", in *Behavioral Ecology*, 29 (6), pp. 1415-1425.

Toger, M., Benenson, I., Wang, Y., Czamanski, D., & Malkinson, D., (2018), "Pigs in Space: an Agent-Based Model of Wild Boar (Sus Scrofa) Movement into Cities", in *Landscape and Urban Planning*, 173, pp. 70-80.

Tønnessen, M., (2009), "Umwelt Transitions: Uexküll and Environmental Change", in *Biosemiotics*, 2, pp. 47-64.

Tønnessen, M., (2020), "Current Human Ecology in the Amazon and beyond: a Multi-Scale Ecosemiotic Approach", in *Biosemiotics*, 13, pp. 89-113.

Uchida, K., Suzuki, K. K., Shimamoto, T., Yanagawa, H., & Koizumi, I., (2019), "Decreased Vigilance or Habituation to Humans? Mechanisms on Increased Boldness in Urban Animals", in *Behavioral Ecology*, 30 (6), pp. 1583-1590.

Von Uexküll, J., (1928), *Theoretische Biologie*, Paetel, Berlin (*Biologia teoretica*, tr. by Guidetti L., Quodlibet, Macerata, 2015).

Von Uexküll, J., (1934): *Streifzüge durch die Umwelten von Tieren und Menschen. Ein Bilderbuch unsichtbarer Welten*, Marina von Uexküll (*Ambienti animali e ambienti umani. Una passeggiata in mondi sconosciuti e invisibili*, tr. by Mazzeo M., Quodlibet, Macerata, 2010).

Von Uexküll, J., (1982), "The Theory of Meaning", in *Semiotica*, 42 (1), pp. 25-82.

Wojnarowski, K., Podobinski, P., Cholewinska, P., Smolinski, J., & Dorobisz, K., (2021), "Impact of Estrogens Present in Environment on Health and Welfare of Animals", in *Animals*, 11 (2152), pp. 1-16.

Zuñiga-Palacios, J., Zunida, I., Castellanos, I., Lara, C., & Sánchez-Rojas, G., (2021), "What Do We Know (and Need to Know) about the Role of Urban Habitats as Ecological Traps? Systematic Review and Meta-Analysis", in *Science of the Total Environment*, 780, pp. 1-11.

Logical predictivism: How to fix use-novelty and vindicate the Copernican Revolution

Vincenzo Crupi
University of Turin
vincenzo.crupi@unito.it

1. The puzzle of early Copernicanism: Epistemic luck *vs*. vindication

From Copernicus himself up to Kepler and Galilei, Copernicans have been "right for the wrong reasons" (Finocchiaro, 2010), because *there were* no epistemically compelling reasons objectively favoring the Copernican position at that stage – a good deal of research in the history and philosophy of science has converged on this claim. In the jargon of contemporary analytic epistemology, the situation of early Copernicans would then be regarded as one of *epistemic luck*. Roughly, epistemic luck characterizes an agent who happens to have a true belief without adequate justification.[1] The precise scope of the *epistemic luck thesis* about early Copernicanism may vary significantly. For our present purposes, it is safe to focus on a version of the thesis which appears particularly sound and popular. According to such version, Copernicanism has been a matter of epistemic luck *at least* from 1543 (the publication of Copernicus's *De Revolutionibus*) up to, say, 1600, namely a moment in which the Copernican allegiance of both Kepler and Galilei is already documented while their own scientific achievements in astronomy were yet to come. Some authors would be happy to say that

[1] On the assumption that Copernicanism is fundamentally correct, the most relevant specification is probably *veritic* (epistemic) luck: "a person *S* is *veritically lucky* in believing that *p* in circumstances *C* iff, given *S*'s evidence for *p*, it is just a matter of luck that *S*'s belief that *p* is true in *C*" (Engel, 2022: 36).

Vincenzo Crupi, "Logical predictivism: How to fix use-novelty and vindicate the Copernican Revolution", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 123-137

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

Copernicanism eventually got to be vindicated with Newton, as it was subsumed under a more comprehensive theory of unrivalled success (e.g., Salmon, 1990: 190). Others might want to insist that heliocentric astronomy remained ultimately unsteady until more "direct" and "physical" evidence of the Earth's motion became available in the XVIII and XIX centuries (see Graney, 2015: ch. 10).

The textual evidence about the popularity of the epistemic luck thesis is sparse but consistent, spanning now more than a century. According to Pierre Duhem's thoughtful discussion in *To Save the Phenomena*, a considered attitude of antirealism fostered by the astronomical tradition led competent observers such as Andreas Osiander and Cardinal Bellarmine to duly appreciate that heliocentric and geocentric systems were empirically on a par at the time, and therefore scientifically on a par too. As Duhem famously and firmly concluded, we are "compelled to acknowledge and proclaim that logic sides with Osiander, Bellarmine, and Urban VIII, not with Kepler and Galilei – that the former had understood the exact scope of the experimental method and that, in this respect, Kepler and Galilei were mistaken" (Duhem, 1908: 113). Fifty years on, another seminal reference is of course Thomas Kuhn. In a key passage of *The Copernican Revolution*, he notes that "each argument" originally put forward by Copernicus "cites an aspect of the appearances that can be explained by *either* the Ptolemaic *or* the Copernican system". The insistence of Copernicus on the greater "harmony" of heliocentrism, Kuhn points out, could only be appealing to a "limited and perhaps irrational subgroup of mathematical astronomers". Only in hindsight can one appreciate that some of them "fortunately" did follow their "Neoplatonic ear" (Kuhn, 1957: 181). And a major theme of Kuhn's view of science is of course that one should strenuously resist turning the benefit of scientific hindsight into a form of hindsight bias in historical matters.

Notably, unlike other implications of Duhem's or Kuhn's work, the epistemic luck thesis about early Copernicanism does not seem to have lost ground over time.[2] As recently as 2011, historian Robert Westman introduced his impressive reconstruction of *The Copernican Question* noting that "Copernicus had opened a question […] which previously had not been seen to possess far-reaching consequences: how to choose between different models of heavenly motion *supported indifferently* by the same observational

[2] Swerdlow (2004: 88) seems to offer a forceful but occasional exception: "There is altogether too much literature today — ultimately, I think, inspired by Duhem and his nonsense about 'saving the phenomena' — that holds that Copernicus had no good reasons to believe his theory to be a true description of the world. He had very good reasons and quite a lot of them."

evidence" (Westman, 2011: 5, emphasis added). Recent extensive work on anti-Copernican astronomy *after* Kepler and Galilei (especially the interesting case of Riccioli, 1651) yielded even stronger claims, if anything. According to Graney, for instance, "in the middle of the seventeenth century […] science backed geocentrism" (Graney, 2015: 144-145; and also see Marcacci, 2015). As for late Twentieth century philosophy of science, Wesley Salmon provides a striking example: "until Newton's dynamics came upon the scene, it seems to me, Thyco's [geostatic] system was clearly the best available theory" (Salmon, 1990: 190). And physicists themselves are apparently no exception: according to Carlo Rovelli, for instance, "Kepler trusted Copernicus' theory *before* its predictions surpassed Ptolemy's" (Rovelli, 2019: 120; also see Timberlake & Wallace, 2019: 144-145).

In the rest of this contribution, I plan to challenge the epistemic luck thesis and argue that, given the information that was actually available in the relevant historical context, it was *not* just a matter of luck that the Copernican view turned out to be correct. It was instead a matter of plausible epistemic justification through sound scientific methodology. Let us call this the *vindication thesis*. My version of the vindication thesis revives Lakatos and Zahar's (1975) view that Copernicus' programme had a remarkable amount of "*immediate* support" from known phenomena that was not matched by the traditional geostatic approach, even if both parties were able to account somehow for all essential facts established in the late Sixteenth century.[3] This will require a revised discussion of the use-novelty of empirical facts in science, which actually amounts to a relatively new tentative approach to the demarcation between empirical success and mere accomodation of known phenomena (see Barnes, 2022, for a valuable survey). The next sections will lay out such proposal and also provide a characterization of the two contenders, namely, Copernicanism and Sixteenth century geocentrism.

[3] Alternative routes to vindication should be mentioned, although I find them ultimately inconclusive. Tipler and Bollinger (2015), for instance, have pursued a rather detailed analysis of empirical accuracy for Ptolemy *vs*. Copernicus against Brahe's quantitative data and "found, on balance, Copernicus to be superior". Others may try to insist on some further discriminating criterion ("simplicity" is of course a major option) as an effective basis to favor Copernicanism as objectively and epistemically superior to its competitors in the relevant time frame (e.g., when Kepler and Galilei decided to join the Copernican camp). See Hall (1970) for an important example of this strategy. Also see Sober (2015: 12-21) for more relevant material on this account.

2. Logical predictivism

Let *S* be a set of empirical findings established by scientific observation and let *T* be a theory (virtually *any* theory) postulating principles, structures, and/or processes underlying the "phenomena" encoded in *S*. As it turns out, it is a crucial fact of the philosophical analysis of science that, as a matter of logic, it will always be possible to derive all elements in *S* as consequences of a "theoretical cohort" integrating *T* with a relevant set of auxiliary assumptions. But this means that an alternative theory *T** could also be aligned with *S* in the same way, namely as embedded in a suitable theoretical cohort.[4] Duhem (1906) is of course a seminal source for this paramount methodological circumstance (see Laudan, 1990: 274, for a more recent statement), which also serves as an undisputed starting point for Lakatos and Zahar. As they say, "any two rival research programmes can be made observationally equivalent by producing observationally equivalent falsifiable versions of the two with the help of suitable *ad hoc* auxiliary hypotheses" (Lakatos & Zahar, 1975: 180).[5] *Duhemian corollary* will work as a convenient shorthand for this statement. Zahar's "new conception" of "novel fact" was meant to go beyond this kind of "uninteresting" empirical equivalence and to specify how the same evidence may still give more support to one theory against another "depending on whether the evidence was, as it were, 'produced' by the theory or explained in an *ad hoc* way". In what follows, much in line with important work by Worrall (2002, 2006), I will employ a minimal implementation of use-novelty which — unlike Zahar's (1973) — squarely avoids reliance on dubious psychological and historical contingencies such as "the reasoning which [the scientist] used to arrive at a new theory" (Zahar, 1973: 219). Consider the following, admittedly basic, characterization of an observable fact *F* as *strongly* confirming a scientific theory *T*:

(a) there exist other observable facts, *E*, such that *F* follows from *T* and *E*; but
(b) *F* does *not* follow from *T* alone; and
(c) *E* and *F* are logically independent.

[4] One such expanded set including theoretical principles and various auxiliary assumptions is sometimes just called a "system". "Theoretical cohort" is a nice terminological variant which I draw from Strevens (2020).
[5] Here, by "observationally equivalent" one should read "such that all *known* observable facts are accounted for by each theory as embedded in its own theoretical cohort".

Each one of clauses (a)-(c) should be meant to apply on the background of further contextually unchallenged assumptions.[6] On this basis, there are two key scenarios in which a researcher will be able to conclude that T is strongly confirmed by F. One amounts to purely temporal novelty: here, the elements in E happen to be already known at a given moment, F is logically derived and *then* established by observation. (In an experimental setting, for instance, the facts in E will typically reflect certain conditions that have been purposely designed and realized in order to check for the occurrence of F, which is expected under those conditions on the basis of T, and ideally not otherwise.) But a situation in which *both* E and F happen to be known is just as much compatible with the fulfilment of (a)-(c), and it arguably captures the idea of so-called *use-novelty*.[7] In Zahar's original cornerstone case, for instance, observationally established facts about the solar system turn out to be sufficient and non-redundant to derive from Einstein's theory of general relativity the already known and otherwise independent fact of Mercury's precessing perihelion and its observable consequences. As all three clauses above are satisfied in this case, evidence about Mercury's perihelion qualifies as an empirical success of the theory regardless of whether Einstein himself may have hoped or even planned to address that problem better than it was handled by classical Newtonian means (see Earman and Janssen, 1993, for a thorough reconstruction). Another related way to look at clauses (a)-(c) is to see them as implying $T \vDash E \supset F$ but ruling out each of $\vDash E \supset F$, $T \vDash E$, and $T \vDash F$. This may be regarded as a situation in which the *connection itself*

[6] The historical evidence in the philosophy of science suggests that a definition of this kind must be liable to charges of triviality. What if T amounts to the combination of $E \supset X$ and $X \supset F$ for arbitrary X and otherwise independent phenomena E and F, for instance? Or what if T combines arbitrary X with the factitious auxiliary $X \supset (E \supset F)$? Here I will not try to develop a formal treatment to neutralize all such frivolous counterexamples (although a subtle potential triviality objection raised by Jason Alexander helped me with the formulation of clause (c)). They will be of no consequence for the subsequent discussion, however. In all cases of interest for us, T will include categorical and unverifiable claims about the world (such as "the Earth revolves around the Sun") that are relevant in the derivation of F from T and E. See Lange's (2004, p. 208) objection to Myrvold (2003) for a related debate.

[7] As far as I can tell, a confirmation theorist who relies on (a)-(c) will elude the troubles raised by Votsis (2014) for "incidental predictivists". Consider the potentially problematic hypothetical case of two scientists A and B such that A derives known fact X from T and known fact Y whereas B derives known fact Y from T and X. If clauses (a)-(c) are satisfied in both cases, my proposal implies that *both X and Y* strongly confirm T. So Votsis's objections do not seem to apply here (Votsis, 2014: 75-76).

between *E* and *F* is made sense of by *T*, not the brute fact of their joint occurrence.[8]

To be sure, this characterization is fully consistent with the Duhemian point that virtually any theory can be tailored and refined to recover known phenomena such as *E* and *F* (see Crupi, 2021), and it is also consistent with the idea that verified observable consequences, even if merely accommodated, can still provide *weak* support for a theory. However, the fact that a key piece of theory (e.g., a Lakatosian hard core, or part thereof) enables the derivation of some of the available evidence from other independent parts of it is arguably contingent on what the theory actually says and is taken as a distinctive element of empirical success. An analogy with evidential reasoning in statistical settings may be helpful. Surely a good measure of fit between, say, a linear model and a relevant data set speaks in favor of a linear interpretation of the underlying process at least to some extent. However, the more stringent demand of so-called *cross-validation* is routinely applied to guard against "overfitting", namely to go beyond the limited support that mere accommodation can provide. If a subset of the data constrains a specification of the model parameters which in turn fares well on a separate subset, the support achieved is taken as clearly stronger (see Schurz, 2014: 92, for a similar remark).

3. A cold case to be revised

An updated account of use-novelty is the first step in my project to recast Lakatos and Zahar's (1975) analysis in a new form, and to counter later criticism, especially by Thomason's (1992). The second step needed is of course a characterization of the theories to be compared. Here, the heliocentric "rough model" or framework (the Lakatosian core of Copernicanism, as it were) will be meant as implying the following claims:[9]

---

[8] As concerns clauses (a)-(c) themselves, I'm really not claiming much originality. In Niiniluoto's (2016) terminology, for instance, the fulfilment of (a)-(c) implies that *T* achieves "deductive systematization" or complies with a "linking up" variant of the notion of "unification" with regards to *E* and *F*. Similar conditions have been also employed to explicate Whewell's celebrated idea of "consilience": see McGrew (2003) and Myrvold (2003). Also see Alai (2014) for a related discussion and proposal.

[9] My reconstruction here is largely consistent with Lakatos and Zahar's (1975) and similar to Carman's (2018). Point (vii), in particular, is explicitly stated early on in the *Commentariolus* as a basic feature of the heliocentric system. In *De Revolutionibus* (Book I, Chapter X), it is presented as the consequence of more fundamental assumptions that are shared through the

- (i) the Sun is stationary;
- (ii) the sphere of the fixed stars, centered (approximately) in the Sun, is at rest;
- (iii) the Earth revolves around the Sun;
- (iv) the Earth rotates around its own axis;
- (v) the Moon orbits the Earth (closely);
- (vi) planets other than the Earth also revolve around the Sun;
- (vii) planets are ordered from the center outward by (strictly) increasing revolution periods.

As concerns the core commitments of the Ptolemaic approach, here is a fitting list for our purposes:

- (i*) the Earth is stationary;
- (ii*) the sphere of the fixed stars revolves around the central Earth;
- (iii*) the Sun revolves around the Earth;
- (iv*) all planets (including the Moon) revolve around the Earth with a combination of (few) circular motions;
- (v*) heavenly bodies are ordered from the outer sphere inward by decreasing overall rotating speed.

Of course, (i*)-(v*) are all consequences of the *full* Ptolemaic theory that was taught in the schools in Copernicus's time including the sophisticated machinery of deferents and epicycles as appropriately specified. Rather crucially, in the current context, Brahe's model itself is nothing but a specification of the core claims (i*)-(v*) above and indeed no more than a variant of the traditional, full Ptolemaic system. In fact, for *any* "planet", the *actual trajectory* postulated by Brahe *around the (stationary) Earth* is demonstrably identical to the corresponding Ptolemaic trajectory. The only caveat is that the Sun is not always further away than Mercury and Venus, but rather at the center of their epicycles. This difference is of course interesting but immaterial for all astronomical evidence available between *De Revolutionibus* and Galilei's discovery of Venus's phases (in 1610), and thus immaterial for our purposes too. Thus, at least in terms of the methodological question about "immediate support" favoring Copernicus's theory, the (post-Lakatosian) reconstruction outlined above thoroughly includes the Tychonic

---

astronomical tradition (also see Lakatos & Zahar, 1975: 185). This elucidation was prompted by a remark from John Worrall.

system as a specific model entailing the pillars of Ptolemaic geocentrism (i*)-(v*).[10]

Let us now put the pieces together, and check the implications.

Fact 1: *Stations and retrogressions are observed for each of Mercury, Venus, Mars, Jupiter, and Saturn*. Thomason (1992) questions that this major point from Lakatos and Zahar (1975: 185) may strongly support the Copernican framework on the grounds that, historically and psychologically, Fact 1 (a "*dominant* problem in Western astronomy", Lakatos & Zahar, 1975: 182) was something that Copernicus definitely *did* want to account for when devising his theory. By our criterion of empirical success (as distinct from accommodation), this is irrelevant, however. Logically, as soon as observational evidence *E* indicates the non-redundant fact of the very existence of a (Copernican) planet (i.e., a major heavenly body other than Moon, Sun, and fixed stars), the Copernican framework (i)-(vii) immediately entails Fact 1 as concerns that object. On the other hand, Fact 1 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 1 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated. As a consequence, Fact 1 does provide strong and immediate support to the Copernican position against the Ptolemaic approach in our revised reconstruction.[11]

[10] One objection here might be that the Tychonic model cannot entail (iv*) just because for Tycho the Sun is the center of simple circular epicycles for each planet. I take this to be an inconsequential semantic issue, however. In both (i)-(vii) and (i*)-(v*), I employ "to revolve" to denote a periodic motion around a stationary center. This is quite consistent with the planets "orbiting" the Sun for Tycho, much as the Moon orbits the Earth for Copernicus. Once this innocent stipulation is clarified, I submit that the traditional Tychonic model does verify (i*)-(v*). (I thank José Díez for pressing me on this point.)

[11] Fact 1 is a qualitative statement. However, in an insightful footnote, Thomason (1992: 181, n. 19) makes a striking observation concerning a more quantitative aspect of these phenomena: in the Copernican approach, the appearance of retrograde motion for superior planets such as Saturn can be large enough to be easily detected only in presence of a "considerable gap" with the fixed stars. The fascinating implication is that, conversely, the observable amplitude of the retrogressions of superior planets may be a basis for a Copernican to infer a large distance of the fixed stars. This in turn would potentially make an empirical success of a fact that no vindicationist seems to have ever dared to classify as more than a (reasonable) accommodation, namely the failed detection of stellar parallax (see, e.g., Worrall, 2002: 198).

Fact 2: *Mercury and Venus are never seen to go in opposition*. Thomason (1992) does not address this point from Lakatos and Zahar (1975: 186), but he could have easily objected that, here again, Fact 2 was an established phenomenon that Copernicus did want to account for when devising his theory. Yet Fact 2 is entailed by the Copernican framework (i)-(vii) along with observational evidence *E* such as a small observed interval between two successive conjunctions (less than a year) for Mercury and Venus, implying the non-redundant statement that both planets are internal. On the other hand, Fact 2 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 2 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated. As a consequence, Fact 2 does provide strong and immediate support to the Copernican position against the Ptolemaic approach in our revised reconstruction.

Fact 3: *Mercury's retrogressions are seen to be more frequent than Venus's*. Thomason addresses a closely related point from Lakatos and Zahar (1975: 186) and questions that it may strongly support the Copernican approach for "it seems plausible to hold that [it] played some role guiding Copernicus to the view that the Sun was in the center of the planets' orbits" (Thomason, 1992: 185). Yet Fact 3 is entailed by the Copernican framework (i)-(vii) along with evidence *E* such as a smaller observed interval between two successive conjunctions for Mercury than for Venus, implying the non-redundant statement that the former must be the innermost internal planet. On the other hand, Fact 3 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 3 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated. As a consequence, Fact 3 does provide strong and immediate support to the Copernican position against the Ptolemaic approach in our revised reconstruction.

Fact 4: *Intervals between successive conjunctions are smaller for Mercury than for Venus.* This point is not addressed by either Lakatos and Zahar (1975) or Thomason (1992), but it is of interest in our perspective. We have seen that observational information about successive conjunctions can complement the Copernican framework (i)-(vii) entailing the ordering of internal planets, by which Fact 3 can then be derived. In addition, this situation is largely symmetric: indeed, Fact 4 is entailed by the Copernican framework (i)-(vii) along with *E* now meant as known observable facts mentioned above. More precisely, because Mercury and Venus are never seen to go in opposition (Fact 2), the theory entails that they must be internal planets, and because retrogressions are seen to be less frequent for Venus than for Mercury (Fact 3), the latter must be the innermost, with a shorter orbital period and thus more frequent conjunctions. On the other hand, Fact 4 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with either *E* or any other independent fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 4 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated.

Fact 5: *The length of Venus's retrograde arc is seen to be greater than Mercury's*. This is a case that Thomason himself allows as use-novel for Copernicus (from *De Revolutionibus*: Book I, Chapter X) because, although of course known, it does "not seem obviously relevant to the structure of the cosmos" (Thomason, 1992: 188), and thus to the guiding explanatory aims of Copernicus' inquiry. In our perspecrive, Fact 5 is entailed by the Copernican framework (i)-(vii) along with observational evidence *E* such as the interval between two successive conjunctions and relevant angular measurements implying a non-redundant assessment of the magnitude and period of Mercury's and Venus's motion as referred to the Sun. On the other hand, Fact 5 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 5 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated.

Fact 6: *Mars, Jupiter, and Saturn are all seen to always retrogress at opposition*. Fact 6 is considered but dismissed by Thomason (1992: 188). Yet Fact 6 is entailed by the Copernican framework (i)-(vii) along with known evidence *E* such as the observation of a quadrature for each of Mars, Jupiter, and Saturn, implying the non-redundant fact that all three planets are external. On the other hand, Fact 6 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 6 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated.

Fact 7: *Jupiter's retrogressions are seen to be more frequent than Mars's, and Saturn's more frequent than Jupiter's*. This point (from *De*

Revolutionibus: Book I, Chapter X) is not addressed by either Lakatos and Zahar (1975) or Thomason (1992). Fact 7 is entailed by the Copernican framework (i)-(vii) along with evidence *E* such as a larger observed interval between two successive conjunctions for Mars than for Jupiter, and for Jupiter than for Saturn (all of which greater than a year), implying the non-redundant statement that Mars must be the innermost external planet, and Saturn the outermost. On the other hand, Fact 7 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 7 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated.

Fact 8: *Intervals between successive conjunctions are smaller for Saturn than for Jupiter, and smaller for Jupiter than for Mars.* This point is not addressed by either Lakatos and Zahar (1975) or Thomason (1992), but it is of interest in our perspective. We have seen that observational information about successive conjunctions can complement the Copernican framework (i)-(vii) entailing the ordering of external planets, by which Fact 7 can then be derived. In addition, this situation is largely symmetric: indeed, Fact 8 is entailed by the Copernican framework (i)-(vii) along with *E* now meant as known observable facts mentioned above. More precisely, because Mars, Jupiter, and Saturn are all seen to go in opposition (Fact 6), the theory entails that they must be external planets, and because retrogressions are seen to be less frequent for Mars than for Jupiter, and for Jupiter than for Saturn (Fact 7), the former must be the innermost and the latter the outermost, with decraeasing orbital periods and thus increasingly frequent conjunctions. On the other hand, Fact 8 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with either *E* or any other independent fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 8 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated.

Fact 9: *The length of Mars' retrograde arc is seen to be greater than Jupiter's*, *which is seen to be greater than Saturn's*. Thomason pairs this with Fact 5 as use-novel for Copernicus (Thomason, 1992: 188). In our perspective, Fact 9 is entailed by the Copernican framework (i)-(vii) along with observational evidence *E* such as the interval between two successive conjunctions and relevant angular measurements implying a non-redundant assessment of the magnitude and period of Mars's, Jupiter's, and Saturn's motion as referred to the Sun. On the other hand, Fact 9 does *not* follow from core Ptolemaic assumptions (i*)-(v*) as conjoined with *E* or any other independent observable fact, so in this case clause (a) is violated. Of course, according to the Duhemian corollary, Fact 9 follows from a full Ptolemaic theoretical cohort (Brahe's is one example), but then clause (b) above is violated.

Concluding remarks

Although surely incomplete, the reconstruction above concerning facts (1)-(9) is sufficient to license a key conclusion for our purposes: according to our characterization of empirical success (which recovers Zahar's original motivation, as illustrated by the Einstein/Mercury example), and despite the uncontested truth of the Duhemian corollary, the Copernican view was indeed "immediately supported" by various known facts which did not support geocentric competitors in the same way.[12] It should be clear – but it's worth emphasizing – that this conclusion relies on a broadly Lakatosian distinction between core *vs*. full models.[13] Again following Lakatos and Zahar, I'm not committed to deny the ("uninteresting") traditional remark that, unlike core models, full models of either strain (heliocentric or geocentric) with all their parameter values specified end up being empirically indistinguishable around 1600 in a relevant sense. In particular, one can see that, for all of them, clause (b) of my criterion of strong support is invariably violated. In a more general vein, logical predictivism seems to fully diclose a remarkable subtlety of evidential support: a fact *F* may appear as no more than an accommodation for a given detailed theory *T*, and yet there may be *T** including a subset of (possibly fundamental) claims from *T* such that *F is* a clear predictive success

[12] One may wonder whether my approach leaves *any* room for strong support in favor of the geocentric position. A fascinating example can be drawn from Carman and Díez (2015: 26-28) and concerns a pattern of phases for a superior planet such as Mars. In our terms, from the observationally established fact that Mars is sometimes found at opposition, one can infer by *either* the heliocentric postulates (i)-(viii) *or* the geocentric postulates (i*)-(v*) the observation of a waxing *vs*. waning gibbous disk before and after opposition, respectively. In this sense, my reconstruction converges with Carman and Díez's (2015) point that a geocentric system does get strong empirical success in a case like this (even if the phenomenon happened to be unobserved before modern times).

[13] In a similar fashion, Myrvold's (2003) assessment of the Copernican controversy relied on the contrast of "a bare-bones Ptolemaic hypothesis with a bare-bones Copernican hypothesis" rather than the corresponding "fully specified models of the heavens, with all parameters filled in". A Lakatosian approach, equipped with a core / programme distinction, can provide a motivation for this move.

of *T**. Arguably, neglect of this circumstance is one key hidden flaw of the popular epistemic luck thesis about the Copernican revolution.

Acknowledgments

This work started from an engaging discussion with Enzo Fano e Flavia Marcacci, whom I want to thank heartily. I'm also grateful for the thoughtful feedback from attendees at the 2023 Conference of the Italian Society for Logic and the Philosophy of Science in Urbino, to Gustavo Cevolani who kindly read and commented an advanced draft, and to José Díez for an extended and sustained exchange of ideas.

References

Alai, M. (2014), "Novel predictions and the no-miracle argument", in *Erkenntnis*, 79: 297-326.

Barnes, E.C. (2022), "Prediction versus accommodation", in E.N. Zalta and U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). URL: https://plato.stanford.edu/archives/win2022/entries/prediction-accommodation.

Carman, C. C. (2018), "The first Copernican was Copernicus: The difference between pre-Copernican and Copernican heliocentrism", in *Archive for the History of Exact Sciences*, 72: 1-20.

Carman, C. C. & Díez, J. (2015), "Did Ptolemy make novel predictions? Launching Ptolemaic astronomy into the scientific realism debate", in *Studies in the History and Philosophy of Science* A, 52: 20-34.

Crupi, V. (2021), "Confirmation", in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition). URL: https://plato.stanford.edu/archives/spr2021/entries/confirmation.

Duhem, P. (1906), *The Aim and Structure of Physical Theory*, Princeton University Press, 1991.

Duhem, P. (1908), *To Save the Phenomena*, University of Chicago Press, 1969.

Earman, J. & Janssen, M. (1993), "Einstein's explanation of the motion of Mercury's perihelion", in J. Earman, M. Janssen, & J. Norton (eds.), *The Attraction of Gravitation: New Studies in the History of General Relativity* (pp. 129-172). Boston: Birkhäuser.

Engel, M. (2022), “Evidence, epistemic luck, reliability, and knowledge”, in *Acta Analytica*, 37: 33-56.

Finocchiaro, M. A. (2010), *Defending Copernicus and Galileo: Critical Reasoning in the Two Affairs*, Dordrecht: Springer.

Graney, C. M. (2015), *Setting Aside All Authority*, University of Notre Dame Press.

Hall, R. (1970), “Kuhn and the Copernican Revolution”, in *British Journal for the Philosophy of Science*, 21: 196-197.

Kuhn, T. (1957). *The Copernican Revolution*. Harvard University Press.

Lakatos, I. & Zahar, E. (1975), “Why did Copernicus’s research programme supersede Ptolemy’s?” in R. Westman (ed.), *The Copernican Achievement* (pp. 354-383). University of California Press, 1975. Reprinted in I. Lakatos, *Philosophical Papers* I: *The Methodology of Scientific Research Programmes* (pp. 168-192). Cambridge University Press, 1978.

Lange, M. (2004). Bayesianism and unification. *Philosophy of Science*, 71: 205-215.

Laudan, L. (1990). Demystifying underdetermination. In C.W. Savage (ed.), *Minnesota Studies in the Philosophy of Science* (pp. 267-297). Minneapolis: University of Minnesota Press.

Marcacci, F. (2015), “The world-system of Giovanni Battista Riccioli and the phases of Venus and Mercury”, in *Advances in Historical Studies*, 4: 106-117.

Margolis, H. (1991), “Tycho’s system and Galileo’s *Dialogue*”, in *Studies in History and Philosophy of Science*, 22: 259-275.

McGrew, T. (2003)., “Confirmation, heuristics, and explanatory reasoning”, in *British Journal for the Philosophy of Science*, 54: 553-567.

Myrvold, W.C. (2003), “A Bayesian account of the virtue of unification”, in *Philosophy of Science*, 70: 399-423.

Niiniluoto, I. (2016), “Unification and confirmation”, in *Theoria* (Spain), 31: 107-123.

Riccioli, G. B. (1651), *Almagestum Novum*, Bologna.

Rovelli, C. (2019), “The dangers of non-empirical confirmation”, in R. Dardashti, R. Dawid, & K. Thébault (eds.), *Why Trust a Theory?*, Cambridge University Press, pp. 120-124.

Salmon, W. C. (1990), “Rationality and objectivity in science, or Thomas Kuhn meets Thomas Bayes”, in C. W. Savage (ed.), *Minnesota Studies in the Philosophy of Science,* Minneapolis: University of Minnesota Press, pp. 175-204.

Schurz, G. (2014), "Bayesian pseudo-confirmation, use-novelty, and genuine confirmation", in *Studies in History and Philosophy of Science*, 45: 87-96.

Sober, E. (2015), *Okcham's Razor: A User's Manual*, Cambridge University Press.

Strevens, M. (2020), *The Knowledge Machine: How Irrationality Created Modern Science*, New York: Northon & Company.

Swerdlow, N. M. (2004), "An essay on Thomas Kuhn's first scientific revolution", in *The Copernican Revolution*. *Proceedings of the American Philosophical Society*, 148: 64-120.

Thomason, N. (1992), "Could Lakatos, even with Zahar's criterion for 'novel fact', evaluate the Copernican research programme?", in *British Journal for the Philosophy of Science*, 43: 161-200.

Timberlake, T. & Wallace, P. (2019), *Finding Our Place in the Solar System: The Scientific Story of the Copernican Revolution*, Cambridge University Press.

Tipler, F. & Bollinger, W. (2015), "Ptolemy *vs*. Copernicus", in *Inference*, 1, n. 3. URL: https://inference-review.com/article/ptolemy-versus-copernicus.

Votsis, I. (2014), "Objectivity in confirmation: Post hoc monsters and novel predictions", in *Studies in History and Philosophy of Science*, 45: 70-78.

Westman, R.S. (2011), *The Copernican Question: Prognostication, Skepticism, and Celestial Order*, University of California Press.

Worrall, J. (2002), "New evidence for old", in J. Wolenski & K. Kijania-Placek (eds.), *In the Scope of Logic, Methodology, and Philosophy of Science,* Dordrecht: Kluwer, pp. 191-209.

Worrall, J. (2006), "Theory-confirmation and history", in C. Cheyne & J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave*, Dordrecht: Kluwer, pp. 31-62.

Zahar, E. (1973), "Why did Einstein's programme supersede Lorentz's?", in *British Journal for the Philosophy of Science*, 24: 95-123 and 223-262. Reprinted in C. Howson (ed.), *Method and Appraisal in the Physical Science,* Cambridge University Press, pp. 211-275.

Mercury’s perihelion anomaly as a use-novel confirmation of general relativity

Vincenzo Fano
University of Urbino
vincenzo.fano@uniurb.it

Introduction

It is well-known that use-novel confirmation is essential in evaluating a scientific theory from an epistemological and a metaphysical point of view[1]. “Metaphysical” in the sense that use-novel prediction increases our confidence that the confirmed theory is true (epistemological value), and then it says at least partially how reality is. Use-novel confirmations are those confirmed predictions of a theory (1) based on evidence that could also come from experiments realized before the formulation of the theory. (2) Still, in any case, this evidence should be inhomogeneous[2] with respect to that used in building the new theory. One of the most famous examples of use-novel confirmation is Einstein’s prediction in 1915 of the anomaly of the precession of Mercury, one of the most critical steps in the final development of general relativity.

This paper aims to investigate the definition of use-novel confirmation proposed by Mario Alai (2014), in view of the famous old evidence confirmation of the anomaly in the precession of Mercury’s perihelion predicted by general relativity. In the next section (2), I briefly discuss the definition of use-novel confirmation, and then (3) I sketch Einstein's (1915b)

[1] One of the first to emphasize this point was Zahar (1973), who used the case of Mercury’s perihelion to exemplify his thesis. In general, see Barnes (2022).

[2] Here, “inhomogeneous” means objects physically characterized in a clearly different way. For instance, in the case of the validity of Newton’s gravitation law, graves on the Earth and planets in the solar system.

Vincenzo Fano, “Mercury’s perihelion anomaly as a use-novel confirmation of general relativity”, in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 139-151

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

calculation. In section 4., I adjust the definition of use-novel confirmation in view of Einstein's reasoning.

1. What is a use-novel confirmation?

The definition proposed by Alai (2014: § 3)[3] is valuable to understand what a use-novel confirmation is[4]. A set of evidence *e* is a use-novel confirmation of *T* iff:

> i) *e* is very improbable unless *T* is true;
> ii) *e* was not used in formulating *T*;
> iii) *e* is a kind of evidence strongly inhomogeneous with respect to the evidence used in formulating *T*.

Each one of these criteria deserves a bit of discussion. The first one is connected to the idea that the less an event is probable, the more informative its occurrence is. Therefore, if evidence *e* without a specific theory *T* is improbable and the theory *T* predicts *e,* and indeed one finds *e*, then *T* is strongly confirmed by *e*. This means that use-novel confirmations should concern facts which do not already have a good explanation.

Concerning the second condition, one should consider that it should not be intended in a historical-psychological sense. ii) means only that from a logico-mathematical point of view, *e* would not be needed in the formulation of *T*. Here, an example could help. A hypothetical physicist, Ga knows that the acceleration of falling bodies on the Earth is proportional to the square of time. Suppose that Ga launches many graves and experimentally s/he discovers that the proportionality constant value is 9.8 m/s^2. Then Ga proposes the new hypothesis that on the Earth, falling bodies follow the law "$s = 9.8t^2$", where *s* is the travelled distance and *t* is the elapsed time. Let *e* = "9.8 m/s^2 be the acceleration of falling bodies on the Earth", and *T* = "$s = 9.8t^2$"; then, we should say that Ga used *e* in formulating *T*. On the other side, let us consider N, who knows that the law of falling bodies on the Earth is $s = 9.8t^2$, and s/he supposes that the law is $s = MG/R^2(t^2)$, where *M* is the mass of the Earth, *G* is Newton's constant of gravitation and *R* the terrestrial radius. Let be *e* = "$s = 9.8t^2$" and *T* = "$s = MG/R^2(t^2)$". In this case, one can say that N does not use *e* in formulating *T*.

[3] Leplin (1997, p. 63) proposes an interesting definition of novelty, which is similar, but different, with respect to Alai's.
[4] Barnes (2022: 8) calls Alai's approach "functional novelty."

Concerning iii), I quote Alai's words: "Although heterogeneity (like similarity) is an intuitively clear notion, it is not easily characterizable since it is gradual and relative. But our criterion can be that a datum is heterogeneous to the essentially used data when it is not inferable from the latter by some standard generalization procedure, without essentially involving the theoretical (unobservable) mechanisms of *T*."

Alai's criteria are good for identifying use-novel confirmations, not for establishing its epistemological value. I do not say much about whether use-novel confirmation is epistemologically relevant and why it is relevant in the affirmative case. The epistemological value of use-novel confirmation is due to its unificatory value. If a certain generalization *L* holds for objects of type *A* and one finds that it holds as well for objects of a completely different kind *B*, this supports that *L* has quite universal validity. However, I will say something more about this in the following.

The question I would like to tackle is: Is the prediction proposed by Einstein on 18 November 1915 of the precession of Mercury's perihelion a use-novel confirmation in Alai's sense?

2. Einstein's calculation

Planets orbiting around the Sun pass through the closest point, i.e., the "perihelion". Due to the perturbation of other planets, the perihelion of each planet does not stand still, but it describes a curve on the plane of rotation of the planet; that is, the perihelion changes orbit after orbit. This phenomenon is called "precession". The precession of Mercury observed by Newcomb (1895), the most reliable at the beginning of the twentieth century, is about 41" every 100 years bigger than what could be explained by the disturbing action of other planets. Note that the angle is measured in a sexagesimal system, that is 41" means less than 3×10^{-3} of an angle of 1 degree in 100 years! A very precise measurement. Seeliger (1906) proposed the most accepted explanation of the phenomenon. He supposes the effect was due to mass distributed in the Mercurial orbit (Earman & Janssen, 1993: 133). Evidence of this mass is the so-called "zodiacal light", spread in certain parts of the sky, a sign of the presence of matter. This hypothesis also has the epistemological advantage[5] that its effect is flexible; therefore, it could be adapted to the possible emergence of new data.

[5] In physics an important part of calculation consists in adapting mathematical models to available data. Therefore, more the models are flexible, easier this calculation is.

Einstein speaks first of Mercury's problem in 1907, in a letter to Conrad Habicht, where he says generically that he is working on a possible explanation of Mercury's problem based on special relativity (Earman & Janssen, 1993: 135). The documentation at our disposal does not give other hints about Einstein and Mercury until the celebrated manuscript written with Besso in 1913[6]. This manuscript is significant since Besso and Einstein established those techniques on the basis of which Einstein, in 1915, realized his famous Perihelion paper. In that period, Einstein was persuaded that building a general covariant theory was impossible. Indeed, in 1913, Einstein and Grossmann published a theory – the so-called "*Entwurf*" (sketch) – which constituted the framework in which he and Besso calculated the anomaly. The final result of this prediction was, unfortunately, wrong. Indeed, at least in Einstein's memory, this failure is one of the reasons why he abandoned the *Entwurf*. Even if this is true, this does not mean that Einstein is using Mercury's perihelion in building his new theory, but only that he considers an important epistemological fact that his new theory could explain the anomaly. We will come back to this issue.

The outlined development seems to justify the following statements:

1. During the building of general relativity, Einstein was perfectly aware of Mercury's problem, and he hoped to explain the anomaly through relativity.
2. Einstein did not use this anomaly to formulate his theory, but he considered explaining the anomaly a critical test for his new theory.

On November 18, 1915, Einstein presented a paper to the Berlin Academy showing that his new theory can solve the problem of Mercury. Einstein solved the question in 7 days. Even Hilbert was impressed by the rapidity of Einstein's calculation. Indeed, Einstein already knew the necessary approximations to calculate Mercury's orbit from his 1913 manuscript, which he realized with Besso, and it was based on the *Entwurf*. However, the 1913 calculation did not fit the experimental result (see Earman, Janssen, 1993).

Today, the textbook calculation of Mercury's perihelion is based on Schwarzschild's solution (1916) of Einstein's equations. This solution was not available for Einstein in 1915. Moreover, in the calculation, Einstein does not use the final equation of general relativity, but "$G_{\mu\nu} = 0$", where "$G_{\mu\nu}$" is

[6] Here it is possible to read a commented edition of this manuscript https://arxiv.org/abs/2111.11238.

the Ricci tensor. Roughly speaking, the Ricci tensor evaluates as space-time volumes change locally with respect to an Euclidean volume. Note that this equation is also valid for the solar system in the definitive framework of general relativity. Around the Sun, where Mercury travels, in the first approximation, there is no matter. Therefore, the stress-energy tensor is null. One can find the details of Einstein's calculation in Earman, Janssen, 1993. Here, I quote only Einstein's assumptions:

> The metric is stationary, time-symmetric, spherically symmetric, and asymptotically Minkowskian.

Together, these assumptions simplify the calculation a lot. Remember that a rank-2 tensor in 4 dimensions – like Ricci's – is a table of 16 numbers that can change at each space-time point. These 16 numbers are not independent due to the symmetries. Therefore, the problem is simpler.

Then, Einstein chooses a simplifying coordinate system and expresses the equation through Christoffel symbols. (The latter represents a so-called "affine-connection"; that is, they establish how to parallel transport vectors). Einstein was eventually persuaded that Christoffel symbols represent the gravitational field.

After this, Einstein calculates the first and second terms of a series expansion of Christoffel symbols[7], establishing the metric around the Sun, which is precisely what one reaches using Schwarzschild's exact solution. Here, it is very important to emphasize that, arriving at this line element, Einstein eventually understood that space needs not to be flat to recover classical mechanics, as he thought before. In other words, obtaining Newtonian orbits is possible even if the space is not flat.

Using the metric, Einstein calculates the equation of motion. Then he arrives to establish that the precession of each orbit is given by "$GM/(a(1 - e^2)c^2)$, where M is the solar mass, G the gravitational constant, a is the semi-major axis of the elliptic orbit of the planet, e the eccentricity of the orbit and c the velocity of light. Remember that the eccentricity e measures how much an orbit differs from a circle. Putting the actual numbers in this formula, Einstein arrives at an anomaly with respect to Newtonian orbits of 43", compatible with the experimental data found by Newcomb.

I report here a quote from Pais (1982: 253), which explains Einstein's mood:

---

[7] Nevertheless, as Earman & Janssen (1993) show, Mercury's precession is a first-order effect of Einstein's equations.

> The first result was that his theory 'explains quantitatively the secular rotation of the orbit of Mercury, discovered by Le Verrier, without the need of any special hypothesis.' This discovery was, I believe, by far the most potent emotional experience in Einstein's scientific life, perhaps in all his life. Nature had spoken to him. He had to be right. 'For a few days, I was beside myself with joyous excitement'. Later, he told Fokker that his discovery had given him palpitations of the heart. What he told de Haas is even more profoundly significant: when he saw that his calculations agreed with the unexplained astronomical observations, he had the feeling that something actually snapped in him.

Einstein's communication caused many reactions in successive years. Perhaps one of the most interesting from an epistemological point of view is that of the mathematician and geophysicist Harold Jeffreys. Jeffreys is well known, above all, for his Bayesian book on probability and statistics (1939); he did not immediately accept Einstein's solution (Jeffreys, 1916), because it was not flexible enough. Indeed, for instance, according to him, if in the future one found that another factor was causing a part of the anomaly, which could explain 10" of the effect at this point, Einstein's result would bring to the wrong datum of 53". On the contrary, a flexible cause, as the zodiacal matter, can be adjusted. Jeffreys will overcome his doubt only in 1919

Before concluding this section on the genesis of the perihelion's paper, one should ask what the experimental and theoretical guides were for Einstein in building his new theory.

Summing up the scholarship extrapolated by Renn (2007), one can say that Einstein's heuristic is based on one side on what we can call "theoretical evidence" and on the other on certain fundamental principles. The term "theoretical evidence" means data that comes from a mere theoretical discussion, even if not observed. Grossly approximating the historical reality, one could say that the main principles are:

A. Einstein's correspondence principle. The new theory should reproduce classical mechanics for weak gravitation.

B. Principle of equivalence. Gravitation is locally equivalent to acceleration – think of the famous lift thought experiment.

C. Generalized relativity principle. The physics should be the same for reference systems uniformly moving along a straight line and for accelerated systems, both rectilinear and circular.

D. Mach's principle, i.e. that masses cause all inertial forces. We know that after the discovery of the expansion of the universe, Einstein completely abandoned this principle.

E. Conservation of energy and momentum.

F. The geometry of spacetime must be non-Euclidean, as shown by the fact that even in special relativity, due to the contraction of length, the ratio between the diameter and the circumference of a rotating disk must be smaller than π.

In the same years, Einstein accepted this theoretical evidence:

a. Special relativity cannot account for gravity because gravity, as an action at a distance, seems speedier than light, and because apparently, in special relativity, Galileo's principle that all bodies fall the same way under gravitational forces seems violated.
b. Since gravity is like acceleration, gravity bends light and slows down clocks.

Each of these items deserves historical discussion; nonetheless, now we have a better idea of the development of general relativity and the role played by the Mercury problem in the genesis of general relativity. The point of this tentative and probably imprecise list is to show that the perihelion anomaly does not play any further[8] logical role in the discovery.

Beginning with the perihelion's paper, it becomes classical to quote three new predictions of general relativity: the bending of light, the redshift of light, and the anomaly in the precession of Mercury's perihelion. The first one to have a clear confirmation, since already in 1895 data were available, is the Mercury's effect.

3. Is Mercury's prediction a use-novel confirmation?

We can now investigate whether Einstein's prediction of Mercury's anomaly in 1915 is a use-novel confirmation in Alai's sense.

[8] "Further" with respect to being a test of the new theory.

Let us consider Einstein's full-fledged equations – even if they were published only in the later paper by Einstein (1915b) – as the theory *T*, and the data made available by Newcomb (1895) as the evidence *e*.

The third criterion proposed by Alai is:

iii) *e* is a kind of strongly inhomogeneous evidence with respect to the evidence used in formulating *T*.

Indeed, this criterion seems trivially satisfied because, as outlined before, in formulating general relativity, Einstein was guided by many principles, A-F, and by two pieces of *theoretical* evidence, a-b, that is, phenomena that are consequences of the equivalence principle and of the gravity force in a special relativity framework. Mercury's anomaly is a piece of experimental evidence. Hence, the inhomogeneity is complete.

The other two criteria are much more problematic. Let us consider the second:

ii) *e* was not used to formulate *T*.

The first answer seems that, in our case, *e* does not satisfy ii) since Einstein considered Mercury's anomaly since 1907 and dedicated to it two important steps of his road to general relativity. Therefore, we cannot say that Einstein did not use *e* in building his new theory.

Renn & Gutfreund (2024: 165) indeed emphasize that after the failure of accounting for Mercury's anomaly by the *Entwurf*, Einstein changed his mind about the *Entwurf,* also pushed by this failure. Therefore, it seems that indeed Einstein *used* the anomaly in building his new theory.

Nevertheless, one can distinguish two senses of the term "to use":[9]

1) A scientist uses_1 certain evidence *e* for building a theory *T* if *e* is used to fine-tune *T*.
2) A scientist uses_2 certain evidence *e* for building a theory *T* if s/he accepts *T* only when it can explain *e*.

One can also dub "use_1" an *intrinsic* use" and "use_2 an *extrinsic* use" of evidence.

Here, a brief epistemological consideration is in order. A standard objection to the epistemological relevance of use-novel confirmation is that a

[9] Alai (2014) had already clarified this point with a different terminology.

historical-psychological fact, as the use (either intrinsic or extrinsic) of evidence in building a theory, cannot have cognitive relevance (Gardner, 1982). As I emphasized in Section 1, Alai's criteria are helpful in establishing whether a certain *e* is or not a use-novel confirmation. Indeed, the epistemological weight of use-novel confirmation should be given only by the logical relation between *T* and *e.* Nevertheless, in concrete science – especially physics – it is very difficult to establish precisely which is the actual logical relation between two issues, as, for instance, general relativity and Mercury's anomaly[10]. For this reason, Alai's criteria are helpful in establishing how much *e* is independent of *T*. This could be said differently. On one side, how concretely Einstein connected Mercury's anomaly to his field equations cannot be directly relevant from the point of view of the logic of justification. On the other side, many arguments employ evidence *e* in deducing a theory *T,* and many do not use *e*. This holds for Mercury's anomaly and general relativity as well. Therefore, the last word on the logical relation between *e* and *T* is almost impossible. For this reason, a concrete analysis of how the construction of *T* is related to *e* is relevant for hinting at the actual relation between *e* and *T*.

Indeed, from our presentation of Einstein's reasoning, it is evident that Einstein's use of the perihelion is extrinsic. There is only one exception. Historians agree that Einstein intended the correspondence principle to require the metric to be flat with weak gravitation. On the contrary, when calculating the orbit of Mercury, he understood that even in the case of weak gravitation, the metric could be non-flat. In this sense, one can say that, at least indirectly, Mercury's anomaly was used at least partly intrinsically in the genesis of general relativity.

We can now pass to the last criterium:

i) *e* is very improbable unless *T* is true.

Again, at first sight, this criterion is not satisfied. At the beginning of the Twentieth Century, Mercury's anomaly was common knowledge, and certainly, this anomaly was not accepted only after Einstein's 1915 calculation. Therefore, if the criterion is formulated well, it is surely not satisfied. Its probability without general relativity ($p(e \wedge \neg GR)$) was already 1. Nevertheless, one should pay attention to the question of *old evidence*.

---

[10] The thesis that use-novel confirmation could be considered as a proxy of a different case-by-case logical relation between *e* and *T* was proposed by Douglas & Magnus (2013) and called "Pluralist Instrumental Predictivism (PIP)".

Indeed, old evidence *e* with respect to a theory *T* is already certain, even if there is not a suitable theory able to explain *e*. In other terms, $p(T/e) = p(T)$.

In literature, Mercury's anomaly is often quoted as a paradigmatic example of old evidence, which, contrary to what appears in a naive Bayesian approach, strongly confirms general relativity[11]. Moreover, approaching Einstein's result in an *ante litteram* Bayesian framework persuaded Jeffreys to refuse Einstein's success, at least for a while.

Among the non-Bayesian attempts, Norton (2021) deserves special attention. He endorses a material theory of induction, which is a theory that adapts good inferential criteria case by case. Concerning Mercury's anomaly, he emphasizes that in 1915, there was only Einstein's solution on the market of the explanations of the effect. This means that a material *modal* fact of this kind held: "If the world were governed by a theory different than general relativity, then Mercury's anomaly would be very improbable" (Norton, 2011). Nevertheless, the historical investigation does not seem to confirm this thesis. Roseveare (1982, p. 2 and 68ss.) emphasizes that astronomers before 1915 accepted Seeliger's solution almost universally. Jeffreys (1916) found yet good motives for preferring Seeliger's solution. Moreover, the notion of a modal fact introduced by Norton is controversial.

The question would deserve a deeper investigation, but what seems highly improbable is neither '$e \wedge \neg T$' nor '$T \rightarrow e$', but that *an algorithm* (*field equations*) *built without considering* e *results almost exactly in* e.

An example could clarify the point. Let us imagine that in a 2-dimensional Euclidean space, there is a point of coordinates (a, b). Let us consider all straight lines passing for the origin described by the equation "$y = kx$)"; k can assume only discrete values from 0 to 100; let us divide the plane in pixels of unitary surface and suppose that a particular line, only 1 of the hundred possible, intersects the pixel occupied by the point (a, b). Moreover, a deterministic algorithm traces one of the 100 possible lines. Furthermore, we can suppose that the a priori probability distribution of the possible lines is uniform; that is, each line has the same probability of being traced. This is because the algorithm applied to choose the line does not use any information about the situation in which it works. In this condition, there is an intrinsic symmetry of the circumstances, which entails that the a priori probability that the line intersects the right pixel is 0.01. Therefore, the fact that the line intersects the right pixel is highly informative about how the line has been traced. Something similar happened in the case of Einstein. Why did the calculation developed by his equation give a result so similar to the

[11] See Eva & Hartmann (2020) for a review and use of Mercury's example.

experimental datum? In a certain sense, the epistemological point is precisely the opposite of what was maintained by Jeffreys: the epistemological strength of Einstein's result seems to stay exactly in the fact that the calculation fits so well with contemporary experimental data. Einstein did not use the anomaly to build relativity and made no special hypothesis in the analysis. Why did his equation arrive exactly at that value? Because *e* use-novel confirms general relativity[12].

To sum up, Mercury's anomaly is substantially a use-novel confirmation in Alai's sense. The criterion of inhomogeneity is fully satisfied. Einstein constrained his new theory by general principles and theoretical evidence, not by experimental evidence, as Mercury's anomaly. Einstein did not intrinsically use Mercury's perihelion in building his new theory but in the fact that he accepted that spacetime could not be flat in the case of weak gravitation. Finally, one cannot say that *e* without *T* was improbable, but that given certain initial conditions and a certain algorithm, the result of the calculation was precisely the same as the experimental evidence. And this fact is very improbable. Indeed, Einstein's inside snapping when calculating the anomaly in 43" was epistemologically well-grounded.

Conflict of interest:
There are no relevant financial or non-financial competing interests to report.

Acknowledgement

I thank Mario Alai, Giovanni Macchia, and four anonymous referees for their very useful comments.

References

Alai, M. (2014), "Novel Predictions and the No Miracle Argument", in *Erkenntnis*, 79, pp. 297-326; doi.org/10.1007/s10670-013-9495-7.

Barnes, E. C. (2022), "Prediction versus Accommodation", in *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.),

---

[12] Indeed, the situation is more nuanced. In physics all calculations are partly driven by the available evidence. And this is true for Einstein's calculation of the anomaly as well. Therefore, the example of the algorithm is not completely adequate.

URL=<https://plato.stanford.edu/archives/win2022/entries/prediction-accommodation/>.

Douglas, H. & Magnus, P. D. (2013), "State of the Field: Why Novel Prediction Matters", in *Studies in the History and Philosophy of Science*, 44, pp. 580–589; DOI 10.1016/j.shpsa.2013.04.001.

Earman, J. & Janssen, M. (1993), "Einstein's Explanation of the Motion of Mercury's Perihelion", in Earman J. & Janssen M., & Norton J.D. (eds), *The Attraction of Gravitation. New Studies in the History of General Relativity*, Boston: Birkhäuser, pp. 129–172.

Einstein, A. & Grossmann, M. (1913), "Entwurf einer verallgemeinerten Relativitätstheorie und einer Theorie der Gravitation", Leipzig: Teubner.

Einstein, A. (1915a), "Erklärung der Perihelbewegung des Merkur aus der allgemeinen Relativitätstheorie", in *Königlich. Preussische Akademie der Wissenschaften* (Berlin). Sitzungsberichte: 831-839.

Einstein, A. (1915b), "Die Feldgleichungen der Gravitation", in Königlich Preußische Akademie der Wissenschaften (Berlin). Sitzungsberichte: 844– 847.

Eva, B. & Hartmann, S. (2020), "On the Origins of Old Evidence", in *Australasian Journal of Philosophy*, 98, pp. 481-494; doi.org/10.1080/00048402.2019.1658210.

Gardner, M. R. (1982), "Predicting Novel Facts", in *British Journal for the Philosophy of Science*, 33, pp. 1–15.

Jeffreys, H. (1916), "The Secular Perturbations of the Four Inner Planets", in *Monthly Notices of the Royal Astronomical Society*, 77, pp. 112–118.

Jeffreys H. (1919), "On the Crucial Test of Einstein's Theory of Gravitation", in *Monthly Notices of the Royal Astronomical Society*, 80, pp. 138–154.

Jeffreys, H. (1939), *Theory of probability*, Clarendon Press, Oxford.

Leplin, J. (1997), *A novel defense of scientific realism*, Oxford University Press, Oxford.

Newcomb, S. (1895), *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*, Washington, D.C.: U.S. Govt. Printing Office.

Norton, J.D. (2011), "History of science and the material theory of induction: Einstein's quanta, mercury's perihelion", in *European Journal for Philosophy of Science*, 1, pp. 3–27; doi.org/10.1007/s13194-010-0001-7.

Norton, J.D. (2021), *The material theory of induction*, https://www.lse.ac.uk/philosophy/blog/2021/11/25/john-d-norton-the-material-theory-of-induction/.

Pais, A. (1982), *'Subtle is the Lord ... ' The Science and Life of Albert Einstein*, Oxford, Oxford University Press.

Renn, J. (2007), *The genesis of general relativity*, Springer, 4 vols.

Renn, J., Gutfreund, H. (2024), *The Einsteinian Revolution: The Historical Roots of His Breakthroughs*, Princeton University Press.

Roseveare, N.T. (1982), *Mercury's perihelion from Le Verrier to Einstein*, Clarendon Press, Oxford.

Schwarzschild, K. (1916), "Über das Gravitationsfeld eines Massenpunktes nach der Einsteinschen Theorie", in *Königlich Preußische Akademie der Wissenschaften* (Berlin), Sitzungsberichte: 189–196.

Seeliger, von H. (1906), "Das Zodiakallicht und die empirischen Glieder in Bewegung der innern Planeten", in *Königlich Bayerische Academie der Wissenschaften* (Munich), Sitzungsberichte 36: 595-622.

Zahar, E. (1973), "Why did Einstein's programme supersede Lorentz's?", in *The British Journal for the Philosophy of Science*, 24, pp. 95-123 and 223-262.

Scientific Realism and Understanding with Deep Learning Models

Giovanni Galli
University of Teramo
ggalli@unite.it

1. Introduction

In the last few years we have witnessed a breakthrough advancement of technologies assisted with artificial intelligence and the AI has gained the attention of almost every scholars dealing with scientific research, industry, healthcare, law, computer science, philosophy of science, ethics and many other disciplines. In particular, deep-learning systems such as Alpha Fold, able to predict with high accuracy the 3d structures of proteins given their sequence of amino acids, as facilitating the way to predict, explore and manipulate proteins. The release by Google Deep Mind of AlphaFold 3 in May 2024 (Abramson et al., 2024), an AI system capable of predicting protein folding with high accuracy, offers a compelling case study for the interplay between scientific understanding, scientific realism, and AI-driven scientific discovery. In this paper I argue that scientific understanding (SU) gained with specific deep-learning models (DLMs), such as AlphaFold's models, has different justifications than SU achieved with theory-driven models and explanations (De Regt, 2015, 2017; Khalifa, 2017). Moreover, SU with AlphaFold's model output is not supported by explanatory information, as the latter. Nevertheless, I submit that the reliability of the model's output depends on a realist conception of the protein's structure prediction. Since deployment realism (DR) is the type of realism focusing on essentiality, it states that it is in virtue of the essential features of the model's output that the link between the target-system and the protein's model justifies the scientific

Giovanni Galli, "Scientific Realism and Understanding with Deep Learning Models", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 153-172

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

understanding involved in such cases. Scientists gain a non-explanatory SU through these models, the success of which can be defined with realist lenses.

Scientific realism is a "positive epistemic attitude"[1] towards the success of science, based on distinctive features of our best scientific theories and models[2] (Chakravartty, 2017). In particular, I explore the predictive accuracy of AlphaFold models as an example of contemporary AI modelling priorities. Realists often adhere to the no miracle argument[3] (NMA), to explain the success of scientific theories. I claim that the success of DLMs as mediators of scientific understanding has to do with realist constraints concerning, in particular, the features deployment realists identify of theoretical constituents (Alai, 2021). Section 2 relates to the architecture of AlphaFold and SU gained with its models. Section 3 presents the debate about deployment realism. In section 4, I argue that deployment realism, perspicuously integrated into model-driven science, suits well for this purpose of recasting scientific understanding with DLMs. Section 5 is about a version of the no miracle argument in AI-driven science and its implications for deployment realism.

For the sake of clarity, I will use DLM or model to refer to the output of AI systems, such as AlphaFold 3. I will use the term AlphaFold 3 (AF3), or AI system to refer to the architecture of the artificial neural networks on which the output depends.

2. AlphaFold and Scientific Understanding

AlphaFold 3 represents a significant advancement in the field of protein structure prediction, building upon its predecessors (Jumper et al., 2021) to achieve even greater accuracy in modelling the three-dimensional shapes of proteins. This success is mainly attributable to the sophisticated architecture of AlphaFold 3 (Abramson et al., 2024), which integrates cutting-edge developments in artificial intelligence, particularly deep learning, with insights from biology and biophysics. The architecture of AlphaFold 3 is designed to address the complex and multifaceted nature of protein folding, a process where a linear sequence of amino acids folds into a specific, stable three-dimensional structure that determines the protein's function.

---

[1] As Chakravartty (2017) defines it.

[2] The priority over the AI models as representations of target-system is here motivated by the shift in scientific research AI-assisted from the theory-driven inquiry to the model-driven one.

[3] See for example Alai (2013; 2016); see also Rowbottom, Peden & Curtis-Trudel's (2024) for a discussion about NMA applied to AI-driven science.

Understanding this folding process has long been a central challenge in biology, given that the number of possible configurations a protein can take is astronomically large, a problem often referred to as the protein folding problem. In this section, I present briefly the architecture of AlphaFold 3 and how AlphaFold system and its models are a way to achieve scientific understanding of the proteins under scrutiny.

2.1. AlphaFold 3 Architecture

At its core, the architecture of AlphaFold 3 employs a deep learning model that can be described as an ensemble of neural networks specifically tailored to capture the intricacies of protein structures. The model is built on a highly refined version of the transformer architecture, a type of neural network originally developed for natural language processing tasks but now adapted to handle the sequential nature of protein sequences (Abramson et al., 2024). The transformer architecture excels at capturing long-range dependencies within sequences, making it particularly well-suited for understanding how different parts of a protein sequence influence each other during the folding process. In AlphaFold 3, this is crucial, as the folding of a protein often depends on interactions between amino acids that are far apart in the linear sequence but come into close proximity in the final folded structure (Wayment-Steele et al., 2024).

In addition to these architectural elements, AlphaFold 3 leverages a vast amount of biological knowledge encoded within its neural networks. This includes data from evolutionary biology, where the model uses multiple sequence alignments (MSAs) to identify conserved regions across different species that are likely to be structurally or functionally important. By incorporating evolutionary information, AlphaFold 3 can make more informed predictions about the likely structure of a protein, even when direct structural data is unavailable. This is particularly useful for predicting the structure of proteins that have not been experimentally resolved, as the evolutionary data can provide clues about the general shape and function of the protein (Abramson et al., 2024).

A further refinement in AlphaFold 3 is the incorporation of a novel geometric module that explicitly models the spatial relationships between different parts of the protein. This geometric module is designed to handle the three-dimensional nature of protein structures, allowing the model to make predictions that are not only accurate in terms of sequence relationships but also in terms of spatial configuration. The geometric module integrates seamlessly with the rest of the architecture, providing a way to translate the sequence-based predictions into a coherent three-dimensional structure that can be compared to experimental data.

The output of AlphaFold 3 is not just a static model of the protein structure but also includes a measure of confidence in each predicted element of the structure. This confidence measure is derived from the model's internal assessment of how well the predicted structure fits with known data and the internal consistency of the predictions (Townshend et al., 2021; Abramson et al., 2024). By providing a confidence score, AlphaFold 3 allows researchers to assess the reliability of the predictions and to focus their experimental efforts on parts of the protein that may require further validation.

The architecture of AlphaFold 3 represents a sophisticated blend of deep learning techniques, evolutionary biology insights, and geometric modelling, all of which are designed to address the complex challenge of protein structure prediction. Through its multi-scale, iterative approach, the use of attention mechanisms, and the incorporation of geometric modelling, AlphaFold 3 achieves a level of accuracy that brings us closer than ever to solving the protein folding problem (Abramson et al., 2024; Jumper et al., 2021).

2.2. AlphaFold's Models and Scientific Understanding

Scientific understanding has recently become a cornerstone notion in philosophy of science (Schurz and Lambert, 1994; De Regt, Eigner, Leonelli, 2009; De Regt and Dieks, 2005; De Regt, 2017; Khalifa, 2017; Lawer, Khalifa and Schech, 2023), often involving the ability to explain, predict, and manipulate phenomena based on theories, models and empirical evidence. The advent of AlphaFold 3 has introduced a new dimension to how scientists gain such understanding in the domain of molecular biology. Traditionally, scientific understanding[4] in this field relied heavily on labour-intensive experimental techniques, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy, to determine protein structures (Jumper et al., 2021). These methods provided direct

[4] I do not define here the specificity of scientific understanding involved in the scientific research with AlphaFold, as it is also attained through the employment of other deep-learning models. The transition from a minimal degree of understanding to a progressively complex degree of understanding necessitates a greater amount of explanatory information. For a discussion about the objectual form of scientific understanding achieved with AlphaFold models, see Schuster (*forthcoming*).

observational data, forming the backbone of our knowledge and understanding of protein folding and function. However, with the introduction of AlphaFold 3, the pathway to scientific understanding has expanded, encompassing not only empirical observation but also computational predictions that can, in some cases, rival the accuracy of experimental methods. Nevertheless, while the scientific understanding achieved in contexts in which researchers do not rely on AI tools is mainly explanatory (De Regt, 2017; Khalifa, 2017), namely it depends on having an explanation for a phenomenon, the scientific understanding obtained through AF models is of different kind, because we lack all the relevant explanatory information concerning protein folding and the relations of each part of the protein with the molecular environment. In this section, I argue that scientists can gain SU with AlphaFold models, without grasping the explanatory framework concerning the kinetics and thermodynamics of protein folding. They understand how the models can be deployed to solve specific problems, but given the state of the art of the system, they do not gain SU of how and why the proteins fold. In such context, predictions (AlphaFold models) are not equivalent to explanations. Indeed, the models do not answer the central why-questions related to protein-folding.

The first way AlphaFold 3 contributes to scientific understanding is by offering highly accurate predictions of protein structures, which can be used to test and refine existing biological models and explanations. The model itself does not offer *prima facie* explanations of why protein folded in such way. Nonetheless, from the representation of the folded protein it is possible to gain SU. Proteins are the workhorses of the cell, involved in virtually every biological process, and their function is intrinsically linked to their three-dimensional shape. Understanding how proteins fold into their functional forms has been a central question in biology[5], and AlphaFold 3's predictions provide relevant insights into this process. When the AI system predicts a protein's structure, it is not merely generating a hypothetical shape; it is leveraging patterns learned from a vast database of known protein structures and sequences to produce a conformation that is likely to exist in nature. Scientists can then compare these predictions to experimentally determined structures, and when the predictions match the empirical data, it reinforces the underlying biological principles encoded in the model. This ability to cross-validate between prediction and experiment enhances our understanding by providing a robust method for verifying theoretical models of protein folding (Mirabello, Wallner, Nystedt 2024).

[5] The folding problem is one of the crucial issue in biophysics (Parke, 2020).

Moreover, AlphaFold 3 facilitates scientific understanding by enabling the exploration of protein structures that are difficult to determine experimentally. Some proteins are challenging to study using traditional methods due to their size, instability, or the conditions required for them to function. In such cases, AlphaFold 3 can predict structures that serve as valuable hypotheses for how these proteins fold and function *in vivo* (Desai, Kantliwala, Vybhavi, Ravi, Patel H. and Patel J., 2024; Campbell, Walden, Walter, Shukla, Beck, Passmore, Xu, 2024). These predictions allow scientists to generate new hypotheses about the roles of specific proteins in biological processes, guiding future experimental work. For instance, understanding the structure of membrane proteins, which are notoriously difficult to crystallize, is crucial for drug development.

Another significant contribution of AlphaFold 3 to scientific understanding is its ability to provide insights into the evolutionary relationships between proteins (Fleming, Magana, Nair and Tsenkov, et al., 2025). Evolutionary biology posits that proteins sharing a common ancestry will have similar structures, even if their sequences have diverged significantly. AlphaFold 3's predictions can reveal these structural similarities, providing evidence for evolutionary theories that suggest conserved folding patterns across different species. This deepens our understanding of how proteins evolve and adapt, offering a structural basis for the functional diversification observed in nature[6]. By predicting the structures of homologous proteins from different organisms, AlphaFold 3 can help scientists trace the evolutionary pathways that have led to the current diversity of life, offering a concrete connection between molecular structure and evolutionary theory.

Moreover, AlphaFold 3 also plays a critical role in enhancing our understanding of protein dynamics and conformational changes (Krokidis, Koumadorakis, Lazaros, Ivantsik, et al., 2025). Proteins are not static entities; they often undergo significant conformational changes to perform their functions. While AlphaFold 3 primarily predicts the most stable conformation of a protein, the insights gained from these predictions can inform our understanding of the range of conformations a protein might adopt. Scientists can use AlphaFold 3's predictions as starting points for molecular dynamics simulations, which explore how proteins move and change shape over time. This integration of static predictions with dynamic simulations helps bridge

[6] AlphaFold 3 has been used also to reconstruct the genealogic trees of viruses' genetic evolution (Callaway, 2024).

the gap between structure and function, offering a more comprehensive understanding of how proteins work in a cellular context.

The interpretability of AlphaFold 3's predictions also contributes to scientific understanding by providing confidence measures that indicate how reliable each aspect of the predicted structure is (Desai, Kantliwala, Vybhavi, Ravi, Patel H. and Patel J., 2024). These confidence scores are not mere byproducts of the computational process; they offer critical information that scientists can use to assess which parts of a prediction are most likely to be correct. This allows researchers to focus their efforts on the most promising areas of a protein's structure when designing experiments or drugs. By indicating the reliability of different structural features, AlphaFold 3 not only provides a static model but also guides scientific inquiry in a more targeted and efficient manner, enhancing the overall process of discovery.

Concerning the conceptual assumptions underlying AlphaFold 3 models' success, we need to explicate the relation between AI models and reality in science as a baseline to analyse the link between models and target-systems. The accurate predictions generated by AlphaFold 3 suggest that the model is capturing essential aspects of the biological reality, even if the model itself is a product of statistical learning rather than a mechanistic understanding of protein folding (Abramson et al., 2024). This challenge traditional views of scientific understanding that emphasize the need for causal-mechanistic explanatory information (De Regt, 2017, Khalifa, 2017). Instead, AlphaFold 3 demonstrates that accurate prediction can be a powerful form of understanding in its own right, even in the absence of a full causal-mechanistic account. This shifts the emphasis from understanding as explanation to understanding as prediction, particularly in complex systems where direct explanations may be elusive.

3. Scientific Realism and Deployment Realism

We have seen that AlphaFold 3 is a powerful tool that helps scientists to understand protein features and functions. Does this understanding rely on realist assumptions? To answer this, it is better to recall the path of scientific realism in general and deployment realism in particular.

Scientific realism is a central position in the philosophy of science that asserts the success of science and the existence of a mind-independent world that science seeks to explain, describe and understand (Smart 1963; Boyd 1983; Devitt 1991; Kitcher, 1993; Kukla 1998; Niiniluoto 1999; Psillos 1999; and Chakravartty 2007). At its core, scientific realism maintains that scientific theories aim to provide true or approximately true descriptions of the world, including both observable phenomena and unobservable entities, such as electrons, gravitational waves, or genes. The fundamental tenets of scientific realism can be summarized as follows: first, the metaphysical claim[7] that the world exists independently of our thoughts, perceptions, or linguistic practices; second, the semantic claim that scientific theories are intended to be approximately true descriptions of the world; and third, the epistemic claim that successful scientific theories, particularly those that have withstood rigorous testing and empirical validation, give us good reasons to believe that the entities and processes they describe actually exist.

A key feature of scientific realism is its commitment to the truth or approximate truth of scientific theories. This commitment is often justified by what is known as the “no miracles” argument (Putnam, 1975; Alai, 2023) which posits that the success of science in producing reliable, accurate predictions and technological advancements would be miraculous if scientific theories were not at least approximately true representations of reality. Scientific realism thus holds that the empirical success of a theory, namely its ability to predict and explain phenomena, is best explained by the theory’s truth or near-truth. This view also implies a belief in the continuity of scientific progress (Bird, 2007; Dellsén 2021, 2023): while theories may evolve or be replaced, there is a continuity of reference in the core terms of successive theories (e.g., “electron” in classical and quantum physics), and thus, the later theories are expected to capture the truth about the entities that earlier theories only approximated.

Another important aspect of scientific realism is its stance on unobservable entities. Unlike empiricist or instrumentalist views (Rowbottom, 2019) which are often sceptical of the existence of unobservable, scientific realism contends that the existence of such entities is justified by the success of the theories that postulate them. Thus, scientific realism encompasses a robust ontology that includes both observable and unobservable entities as real components of the world.

In contrast to the broad framework of traditional scientific realism, deployment realism (DR) is a more focused and pragmatic variant claiming that a hypothesis is most probably true when it is deployed essentially in a novel prediction (Alai, 2021). I advance that we can extend DR not only to successful theories but also to accurate models. So, DR would emphasise the reliability of scientific models and theories as they are deployed also in

---

[7] For a detailed scrutiny of the relation between scientific realism and metaphysical antirealism, see Alai (2023).

practical applications. Deployment realists share the fundamental realist commitment to a mind-independent world but distinguishes itself by concentrating on the success of scientific theories and models in specific contexts of deployment. Rather than asserting that all aspects of a successful theory and model are true, deployment realism suggests that we have good reasons to believe in the reality of those entities and processes that are directly involved in the successful deployment of the theory and models in practical, often technological, contexts. DR concerning DLMs is motivated by these observations:

1) Scientific DLMs achieve remarkable predictive success.
2) This predictive success is not satisfactorily explained by anti-realist account of science.
3) The models' constituents that are deployed in, or responsible for this predictive success, are often retained from one model to another, even when the starting models are superseded.

The first characteristic of DR is its emphasis on the context of models use. Deployment realism is less concerned with the global truth of an entire scientific theory or model and more with the local truth of the specific components of a theory that are involved in successful applications. For example, in the case of a computational model used in climate prediction, deployment realism would assert the reality of the climate processes that the model successfully captures and predicts, while remaining agnostic or noncommittal about the truth of other aspects of the underlying climate theory that are not directly implicated in the model's success.

A second characteristic of deployment realism is its focus on the epistemic warrant provided by practical success. Deployment realism argues that when a scientific model or theory is successfully deployed in a practical context, such as in technology, medicine, or engineering, this success provides strong epistemic grounds for believing in the reality of the entities and processes that the model or theory posits. This focus on practical success as an epistemic warrant is a key departure from traditional scientific realism, which often emphasizes the theoretical virtues of coherence, simplicity, or explanatory power as indicators of truth.

A third characteristic, and mostly important for DLMs, of DR is its flexibility regarding the opacity of models. Many modern scientific models, particularly in fields like AI and computational science, are complex and opaque, meaning that their internal workings are not fully understood or interpretable by humans. Due to the opacity of DLMs, we do not have any relevant epistemic information about the justification process by which the system gives the output. Traditional scientific realism, with its emphasis on explanation and understanding, may find such models problematic. However, deployment realism accommodates this opacity by focusing on the practical outcomes of deploying these models. It argues that, even if we do not fully understand how a model works, the success of its deployment in real-world contexts can still justify belief in the reality of the entities it models, i.e. the predicted strutures made of each molecular part. Put in that way, in non AI-driven research contexts SU is given by explanations, such as in an AI-driven research context as the case with AF3 models SU is achieved by predictions, which are the the model representations of proteins' structures.

Furthermore, DR acknowledges the provisional nature of scientific knowledge, aligning itself with a more pragmatic and context-sensitive understanding of scientific progress. In this view, scientific knowledge is not seen as a linear progression towards a final, ultimate truth but rather as a series of successful applications that provide increasingly reliable knowledge about specific aspects of the world. This aligns with the modern scientific practice, where theories and models are often revised, improved, or replaced as new data and technologies emerge. Deployment realism thus allows for a dynamic, evolving understanding of scientific knowledge, where the reality of certain entities is continually reinforced through successful deployment, even as broader theories may change.

In defending an application of DR to AI-driven scientific research contexts, such as using AF's models, I am particularly interested in the resulting epistemic trade-offs. While DR capture the models' reliability (in intervention and manipulation), applicability (across contexts), and stability (across theory change), we must be aware that if we are interested in catching the virtues of the whole theories about, for exampla, protein folding, we should look elsewhere.

In sum, DR can be seen as a refined and context-sensitive version of scientific realism, one that emphasizes the practical success of scientific models and theories in specific contexts of application. While traditional scientific realism is concerned with the truth or approximate truth of entire theories and the existence of unobservable entities, deployment realism is more focused on the reliability and epistemic warrant provided by the successful deployment of scientific models.

4. Scientific Understanding and Deployment Realism

The success of AF in accurately predicting protein structures, often at a level comparable to experimental methods, underscores the potential of such models as fundamental components of scientific practice, thereby enhancing our understanding of biological systems. However, we must acknowledge the inherent limitations of AF systems and models in fully capturing the complexity of biological systems.

While some accounts of scientific understanding might rely on instrumentalist and antirealist stances (De Regt, 2017), the kind of scientific understanding addressed here still holds continuity with scientific realist tenets[8]. Scientific understanding, indeed, involves grasping how the entities and processes posited by scientific theories and models relate to the phenomena we observe. Understanding, in this sense, can be defined as having accurate descriptions or predictions, and could also involve the ability to explain and manipulate the natural world based on these scientific representations. The deployment realist argues that models like those provided by AlphaFold are integral to this understanding. They allow scientists to visualize and manipulate the structures of proteins, leading to insights into their functions, interactions, and roles in biological systems, even if they do not provide explanatory information about the specific biophysical causal chain that makes the proteins fold in such and such a way. This practical utility suggests that models are not mere tools of convenience but are essential to the epistemic aims of science, namely, to understand phenomena.

AlphaFold's predictions are a prime example of how models can lead to a non-explanatory scientific understanding. Proteins are complex macromolecules that perform a vast array of functions within living organisms, and their functions are intimately tied to their three-dimensional structures. Traditionally, determining these structures required labour-intensive experimental techniques like X-ray crystallography or cryo-electron microscopy. AlphaFold, however, uses machine learning to predict these structures with remarkable accuracy, providing scientists with a powerful new tool for exploring protein function. These predictions are not merely hypothetical constructs; they can be tested and verified against experimental data, and they often provide insights that were previously inaccessible. This predictive power exemplifies how models can extend our understanding by providing representations that are both accurate and informative, aligning

[8] For a realist reading of scientific understanding, see Galli (2024).

well with the deployment realist's emphasis on the practical success of scientific models.

The success of AlphaFold also raises interesting questions for the broader debate about scientific realism. One of the key challenges for scientific realism has been the so-called "pessimistic meta-induction", which argues that because many successful scientific theories of the past have later been shown to be false, we have reason to doubt the truth of current theories. DR, however, sidesteps this issue by focusing on the parts of science that continue to be successful even as theories change. In the case of AlphaFold, even if future developments in biology or machine learning lead to new models or theories about protein structure, the current success of AlphaFold's predictions demonstrates that the models it generates have a strong claim to be considered at least accurate, if "approximately true" sounds too strong, in their depiction of protein structures. This success supports the deployment realist's claim that models, like theories, can provide genuine understanding of the natural world.

Moreover, AlphaFold highlights the epistemic significance of models in a way that challenges more traditional, theory-centric views of scientific realism. If scientific realism is primarily concerned with the truth of theories, it might overlook the fact that models often play a more direct role in scientific practice. For instance, while the underlying theories of protein folding are important, it is AlphaFold's model, one of the concrete, operational tool, that provides the actionable insights that biologists rely on.

This broader perspective on scientific realism also has implications for how we think about the nature of scientific progress. If models like AlphaFold's are central to scientific understanding, then scientific progress can be seen not just as a matter of developing better theories, but also as a matter of developing better models[9]. The iterative improvement of models, their increasing accuracy, and their expanding applicability all contribute to the advancement of science. This view aligns well with the deployment realist's focus on the practical success of scientific models, suggesting that progress in science is not just about getting closer to the truth, but about developing tools and models that allow us to better understand, predict, and manipulate the world.

[9] Under this light, the application of DLMs to models is consistent with Dellsén's (2016, 2021) view about progress in science.

The connections between scientific realism and scientific understanding is then complex and not wholly analysed yet[10]. However, the growing recognition of the role of models in scientific practice, exemplified by tools like AlphaFold, suggests that models are just as important as theories in fostering scientific understanding.

4.1. Deployment Realism and AlphaFold Deep-learning Models

Deep-learning models, particularly in the domain of bioinformatics and computational biology, have rapidly advanced, achieving significant success in tasks that were previously thought to be intractable.

From the perspective of DR, the success of AlphaFold can be seen as providing a strong case for the reality of the entities and processes the system models, i.e. what the system predicts. We can extend similar claims also to the case of Digital Twins models or Foundational models, which aim to represent a virtual *replica* of a physical phenomenon, setting or environment (Haag and Reiner, 2018). Specifically, DR implies that the accurate predictions made by AlphaFold about protein structures indicate that the underlying biochemical processes it simulates have a basis in reality, as also the proteins' constituents. The fact that these predictions can be verified experimentally, by comparing AlphaFold's predicted structures with those determined through empirical methods such as X-ray crystallography or cryo-electron microscopy, further strengthens the deployment realist's position.

However, the application of DR to DLMs like AlphaFold's raises crucial questions. One key issue is the opacity of these models, the fact that they often operate as "black boxes", with their internal workings being difficult to interpret even by the experts who design them. This opacity[11] challenges traditional notions of scientific understanding, which typically emphasize the importance of having transparent, explanatory models. In the case of AF, while the predictions are highly accurate, the underlying mechanisms by which it arrives at these predictions are not fully understood. DR, therefore, must grapple with the question of whether the success of a system like AlphaFold justifies belief in the reality of the entities it models, even if we lack a clear understanding of how the model works. Still, if the models are

[10] For a debate about interconnected issues between both themes, see Part II of *Scientific Understanding and Representation. Modeling in the Physical Sciences*, edited by Lawler, Khalifa and Shech (2023), pp. 133-199.

[11] For a detailed analysis of different kinds of opacity involved in this context, see Termine and Facchini (2022).

epistemic reliable, it is in virtue of their ability to represent correctly some essential parts of the target-system, as DR requires. Moreover, researchers should be able to apply the model to different contexts, if they are accurate. Furthermore, accurate AF's model, since they are built without explanatory information related to the relevant theories, are also stable across theory change; and this is an important epistemic advantage in comparison to the previous protein models obtained *via* experimental methods.

Testing DR with the case of AF's models reveals that, if there is consensus about the claim that researchers using AF's models gain scientific understanding of the proteins they study, SU can be achieved also without having explanatory information about the relevant phenomena, as in AI-driven science, still having a prediction, namely AF's model, of a phenomenon (protein folded) and thanks to its deployment in the experimental research settings. Even without complete explanations for the predicted structure of the proteins folded, researchers gain SU *via* models generated by an AI system, such as AlphaFold. In absence of explanations, AF's models are a viable epistemic tool to scientific understanding in virtue of the representational properties of the corresponding target-system in the world, namely the molecular protein structures.

While traditional accounts of scientific realism emphasize explanatory understanding, deployment realism suggests that the success of a model in practical cases can itself provide grounds for belief in the reality of the entities and processes it models. In the case of AF, this perspective allows us to view the model's success as indicative of the reality of the molecular structures and interactions it predicts, even if our understanding of how the model works is incomplete.

5. No Miracles Argument in AI and Deployment Realism

The No Miracles Argument (NMA) has long been a cornerstone of scientific realism, positing that the best explanation for the success of scientific theories is that they are at least approximately true (Puntam, 1975; Alai, 2023). Realists argue that the empirical success of a theory, its ability to generate accurate predictions and explain phenomena, would be miraculous if the theory were not at least partially true. This argument underpins the realist conviction that science progressively uncovers truths about the world, even if our theories are imperfect or incomplete. With the advent of AI, particularly in the case of DLMs, the application of the NMA has taken on new significance and complexity. The emergence of NMAAI (Non-Miracles

Argument for AI) requires a redefinition of deployment realism, especially in the context of AI-driven scientific discovery (Rowbottom, Peden, Curtis-Trudel, 2024).

AI systems, especially those leveraging deep learning, present a unique challenge to traditional notions of scientific realism because they operate in ways that are fundamentally different from human-driven scientific inquiry. Unlike traditional scientific theories, which are often grounded in a causal understanding of the phenomena they explain, AI systems like AlphaFold generate successful predictions through sophisticated pattern recognition, rather than through an understanding of underlying mechanisms. The success of these models in generating accurate predictions invites the application of the NMA. According to the NMAAI, the success of AI models like AlphaFold suggests that these models must be tapping into real features of the world, even if they do so in ways that are opaque or inscrutable to human understanding.

Nevertheless, applying the NMA to AI demands reconsidering what it means for a model to be "true" or to represent reality accurately. Traditional scientific realism is concerned with the truth of scientific theories in a representational sense, whether the entities posited by the theory correspond to actual entities in the world, and whether the theory accurately describes the causal mechanisms that produce observable phenomena. AI models, however, do not necessarily offer such representations. Instead, they provide highly accurate predictions based on learned correlations within large datasets. The mechanisms by which these predictions are made often remain hidden within the "black box" of the AI, raising questions about whether these models can be said to possess "understanding" in any meaningful sense, and whether their success can be taken as evidence of their truth in the realist sense (Páez, 2019; Durán, 2021; Sullivan, 2022; Räz and Beisbart, 2022).

Moreover, when applied to AI, deployment realism must be redefined to accommodate the fact that AI models may achieve practical success without offering any explanatory, may it be causal or mechanistic, insight. In the case of NMAAI, deployment realism would not necessarily require that AI models offer true representations of the world in a traditional sense but would instead focus on the reliability and accuracy of the model's predictions as evidence of their approximate truth.

This redefinition of deployment realism in the context of NMAAI involves several considerations. Rather than asking whether a model describes the true underlying mechanisms of phenomena, NMAAI suggests that success in prediction is sufficient to attribute some form of truthlikeness to the model, given by, specifically, a structural correspondence between the model's outputs and the patterns in the world (Galli, 2023). For example, AlphaFold's ability to predict 3D protein structures with high accuracy implies that its internal architecture captures real-world regularities, even if those regularities remain opaque to human understanding. The structural correspondence between the AI-generated models and the actual 3D configurations of proteins suggests that the AI is tapping into real patterns or structures in the world, even if it does not represent these structures in a way that corresponds to human understanding. In this sense, the deployment realism associated with NMAAI would be concerned with the model's ability to reliably generate accurate predictions, rather than its ability to provide causal or mechanistic explanations. This perspective, however, invites several objections. First, one might argue that predictive success without explanatory insight is not sufficient for a realist commitment. Without understanding why a model works, critics contend, we cannot rule out that its success is merely coincidental or the result of overfitting to data. In response, defenders of NMAAI can point out that the sustained and generalizable predictive success of models like AlphaFold's, across varied and previously unseen protein sequences, makes the coincidence hypothesis implausible. Furthermore, while overfitting remains a risk in any statistical model, rigorous testing on novel data provides empirical safeguards that support the reliability of AI-driven predictions. A second objection concerns epistemic opacity: the idea that if we cannot interpret or reconstruct the model's decision-making process, then we cannot claim to "understand" the phenomena it models in any meaningful sense. In this case, deployment realism might appear to collapse into mere instrumentalism. However, a revised form of deployment realism can resist this reduction by appealing to a broader, non-explanatory notion of scientific understanding, one grounded in practical utility and predictive coherence. If scientific realism is to remain viable in the context of AI, it must accommodate forms of understanding that emerge from effective interaction with phenomena, even in the absence of transparent causal explanation (Durán, 2021; Sullivan, 2022; Räz and Beisbart, 2022). A third and more constructive aspect of this redefinition concerns the collaborative nature of understanding in AI-assisted science. AI systems alone do not interpret or contextualize their outputs. It is through human interpretation that the predictive outputs of models like AlphaFold are integrated into broader scientific theories and experimental practices. Thus, deployment realism in the context of NMAAI is not an endorsement of algorithmic autonomy, but a recognition of a hybrid epistemic framework where machine-generated predictions and human interpretative practices co-constitute scientific understanding.

In this new AI-driven landscape, realism is not abandoned but reoriented. The success of AlphaFold's models, provides evidence not for the truth of their internal representations in a semantic sense, but for their capacity to track and exploit real-world patterns in ways that serve epistemic and practical ends. This redefinition of deployment realism aligns with a broader noetic perspective towards model-based, tool-oriented epistemologies suited for the complexities of AI-driven scientific inquiry.

6. Conclusion

The redefined deployment realism, as explored in this paper, shifts the focus from theories to models and from causal explanation to structural correspondence and predictive success in AI-driven science. AI models like AlphaFold's may not represent the causal processes of protein folding in a way that is intelligible to human scientists, but their ability to consistently generate accurate predictions suggests that they are capturing something real about the world. This form of realism does not require that AI models provide a true depiction of the world in the traditional sense but instead argues that their success is evidence of their structural alignment with reality, even if that alignment is not fully understood by human observers.

Furthermore, the redefinition of DR in the context of NMAAI must account for the epistemic opacity of AI models. The fact that these models can generate highly accurate predictions without providing transparent explanations challenges the traditional view that understanding in science is necessarily linked to explanation. Instead, NMAAI suggests that understanding can also be achieved through the reliable generation of predictions, even in the absence of mechanistic insight. This has significant implications for our conception of scientific inquiry, suggesting that it may be possible to gain a form of scientific understanding that is grounded in predictive success rather than explanatory depth.

This form of realism, grounded in the predictive success of AI models, suggests that scientific understanding can be achieved through structural correspondence and reliability, even in the absence of causal explanation. As AI continues to play an increasingly central role in scientific discovery, this form of realism will be crucial for navigating the complex interplay between human understanding and machine-generated insights, ultimately reshaping our conception of what it means to understand the world from both a human and machine perspective.

References

Abramson, J., Adler, J., Dunger, J. *et al*. (2024), "Accurate structure prediction of biomolecular interactions with AlphaFold 3", in *Nature* 630, 493-500.

Alai, M. (2014), "Novel Predictions and the No Miracle Argument", in *Erkenntnis*, 79, 297-326.

Alai, M. (2016), "The No Miracle Argument and Strong Predictivism Versus Barne", in Magnani, L., Casadio, C. (eds) *Model-Based Reasoning in Science and Technology. Studies in Applied Philosophy, Epistemology and Rational Ethics*, vol 27, Cham, Springer, pp. 541-556.

Alai, M. (2023), "Scientific Realism, Metaphysical Antirealism and the No Miracle Arguments", in *Found Sci* 28, 377-400.

Bird, A. (2007), "What Is Scientific Progress?", in *Noûs*, 41, 1, 64-89.

Boyd, R. N. (1983), "On the Current Status of the Issue of Scientific Realism", in *Erkenntnis*, 19(1/3), 45-90.

Callaway, E. (2024), "Where did viruses come from? AI is finding answers", in *Nature*, 633.

Campbell, A. E., Walden, H., Walter, J. C., Shukla, A. K., Beck, M., Passmore, L. A., & Xu, H. E. (2024), "AlphaFold: Research accelerator and hypothesis generator", *Molecular Cell*, Volume 84, Issue 3, 404-408.

Chakravartty, A. (2007), *A Metaphysics for Scientific Realism: Knowing the Unobservable*, Cambridge: Cambridge University Press.

Chakravartty, A. (2017), "Scientific Realism", in *The Stanford Encyclopedia of Philosophy* (Summer 2017 Edition), Edward N. Zalta (ed. by), URL = https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/ (last view 20/09/2024).

De Regt, H. W., Leonelli, S. & Eigner, K. (eds.) (2009), *Scientific Understanding: Philosophical Perspectives*, Pittsburgh, University of Pittsburgh Press.

De Regt, H., W., & Dieks, D. (2005), "A Contextual Approach to Scientific Understanding", in *Synthese*, 144, 137-170.

De Regt, H., W. (2017), *Understanding Scientific Understanding*, Oxford, Oxford University Press.

Dellsén, F. (2016), "Scientific progress: Knowledge versus understanding", in *Studies in History and Philosophy of Science*, 56, 72-83.

Dellsén, F. (2021), "Understanding scientific progress: the noetic account", in *Synthese* 199, 11249-11278.

Dellsén, F. (2023), "Scientific progress: By-whom or for-whom?", in *Studies in History and Philosophy of Science*, Volume 97, 20-28.

Desai, D., Kantliwala, S. V., Vybhavi, J., Ravi, R., Patel, H., & Patel, J. (2024), "Review of AlphaFold 3: Transformative Advances in Drug Design and Therapeutics", in *Cureus*, 16(7), e63646.

Devitt, M. (1991), *Realism and Truth*, Oxford, Blackwell.

Durán, J. M. (2021), "Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare", in *Artificial Intelligence*, 297, 1-14.

Facchini, A., & Termine, A. (2022), "Towards a Taxonomy for the Opacity of AI Systems", in Müller, V.C. (eds), *Philosophy and Theory of Artificial Intelligence 2021*, PTAI 2021, *Studies in Applied Philosophy, Epistemology and Rational Ethics*, vol 63. Cham, Springer.

Fleming, J., Magana, P., Nair, S., Tsenkov, M., *et al*. (2025), "AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities", in *Journal of Molecular Biology*, 2025, 168967.

Galli, G. (2023), "Structure Representation of Deep-Learning Models: The Case of AlphaFold", in *Argumenta*, 9, 1, 43-60.

Galli, G. (2024), "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.

Haag, S., & Anderl, R. (2018), "Digital twin – Proof of concept", in *Manufacturing Letters. Industry 4.0 and Smart Manufacturing*, 15, 64–66.

Jumper, J., Evans, R., Pritzel, A. *et al.* (2021), "Highly accurate protein structure prediction with AlphaFold", in *Nature*, 596, 583-589.

Khalifa, K. (2017), *Understanding, Explanation, and Scientific Knowledge*, Cambridge, Cambridge University Press.

Khalifa, K., Lawler, I. & Shech E. (2023), *Scientific Understanding and Representation: Modeling in the Physical Sciences*, London, Routledge.

Kitcher, P., 1993, *The Advancement of Science*, New York, Oxford University Press.

Krokidis, M. G., Koumadorakis, D. E., Lazaros, K., Ivantsik, O., Exarchos, T. P., Vrahatis, A. G., Kotsiantis, S., &Vlamos, P. (2025), "AlphaFold3: An Overview of Applications and Performance Insights", in *International Journal of Molecular Sciences*, 26(8), 3671.

Kukla, A., (1998), *Studies in Scientific Realism*, Oxford, Oxford University Press.

Mirabello, C., Wallner, B., Nystedt, B. *et al.* (2024), "Unmasking AlphaFold to integrate experiments and predictions in multimeric complexes", in *Nat Commun*, 15, 8724.

Lawler, I., Khalifa, K., & Shech, E. (2023), *Scientific Understanding and Representation*. *Modeling in the Physical Sciences*, New York, Routledge.

Niiniluoto, I. (1999), *Critical Scientific Realism*, Oxford, Oxford University Press.

Páez, A. (2019), "The pragmatic turn in explainable artificial intelligence (XAI)", in *Minds and Machines*, 29, 441-459.

Parke, W., C. (2020), *Biophysics. A Student's Guide to the Physics of the Life Sciences and Medicine*, Cham, Springer.

Psillos, S. (1999), *Scientific Realism: How Science Tracks Truth*, London-New York, Routledge.

Putnam, H. (1975), "What is Mathematical Truth?" in H. Putnam, *Mathematics, Matter and Method. Philosophical Papers Volume I,* Cambridge University Press, Cambridge, pp. 60-78.

Räz, T., & Beisbart, C. (2022), "The Importance of Understanding Deep Learning", in *Erkenntnis*, 1-18.

Rowbottom, D. P. (2019), *The Instrument of Science. Scientific Anti-Realism Revitalised*, London, Routledge.

Rowbottom, D. P., Peden, W., & Curtis-Trudel, A. (2024), "Does the no miracles argument apply to AI?", in *Synthese* 203, 173.

Schurz, G., and Lambert, K., 1994, "Outline of a theory of scientific understanding", *Synthese*, 101, 1, 65-120.

Schuster, A. (2025), (*manuscript*), "Understanding protein folding with Machine Learning Models? The case of Alpha Fold".

Smart, J. J. C. (1963), *Philosophy and Scientific Realism*, London, Routledge & Kegan Paul 1963.

Sullivan, E. (2022), "Understanding from Machine Learning Models", in *Br. J. Philos. Sci.*, 73, 1, 109-133.

Townshend, R. J. L. *et al.* (2021), "Geometric deep learning of RNA structure", in *Science*, 373, 1047-1051.

Wayment-Steele, H. K. *et al*. (2024), "Predicting multiple conformations via sequence clustering and AlphaFold2", in *Nature*, 625, 832-839.

Whitehead's Relational Special Relativity
A Natural Philosophy of Time

Enrico R. A. C. Giannetto
University of Bergamo
enrico.giannetto@unibg.it

1 Introduction

Alfred North Whitehead was born in 1861 and died in 1947. He is very well known as the author, with Bertrand Russell, of the great logico- mathematical treatise *Principia Mathematica* (Whitehead & Russell, (1910-13) and for other mathematical and philosophical works (Whitehead, (1906), (1907), (1929), and Schilpp, (1941), and Northrop, (1941)). Regarding physics, Whitehead is practically known only among few general relativistic theorists for his sort of special-relativistic theory of gravitation, formulated in opposition to general relativity (Whitehead, (1922); Schild, (1956); Synge, (1956); North, (1965) 186-197; Grünbaum, (1973) 48-65, 425-28).

However, in my opinion, the greatest work of Whitehead concerns his physics, even if it has very important philosophical and mathematical implications. And it is not his theory of gravitation, but it is his relational formulation of special relativity, that is completely independent of his gravitational theory and also of his trials to link the world of experience and of perceptual representation and the world of physics at a foundational level (Whitehead, (1920), (1929); Russell, (1927)).

Here, I shall consider his work on special relativity as a time-theory as independent from his special relativistic gravitational theory. His special relativistic gravitational theory is an alternative to Einstein's general relativistic gravitational theory, but his analysis of special relativity in terms of temporal relations among events can be considered a different, relational, interpretation of special relativity.

Enrico R. A. C. Giannetto, "Whitehead's Relational Special Relativity. A Natural Philosophy of Time", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 173-196

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

Here, I shall not consider the temporal evolution of Whitehead’s philosophy till his process cosmological “metaphysics” (Whitehead, (1929). I shall focus my analysis only on his relational interpretation of special relativity in its mathematical formulation: here, Whitehead gave a very fundamental role to time. Spatial mathematical entities (as points, straight lines, etc.) are defined in terms of temporal events and different time reference frames (for example, a straight line in a reference frame at rest is considered in terms of a spatial point as viewed by a reference frame in rectilinear uniform motion – and a spatial point is a series of events). Here, space is space of a plurality of times of becoming.

In my opinion, Whitehead’s process conception of Nature has its roots in this relational interpretation of special relativity where spatial concepts and material bodies are not fundamental entities but are defined in terms of temporal events: here, time is the stuff of which is made reality and it is not reduced to space or to a timeless, eternal four-dimensional space-time.

Only few scientists have in some way developed his perspective on special relativity (Russell, (1914), (1927)). Indeed, Whitehead has given a solution to Ockham, Al Ghazali and Kalam school, and Leibniz' major problem of constructing a relational theory of space, time and motion, and so of geometry (Ockham, (1985); Leibniz, (1849), (1875); Alexander, (1984); Poincaré, (1880); Russell, (1900), (1903); Jammer, (1954), (1957); Huygens, (1905); Korteweg & Schouten, (1920); Reichenbach, (1924); Earman, (1989); Giannetto, (1987), (1995), (2005)), by defining all the fundamental concepts and formulating (special) relativity in terms of event-particle relations (Whitehead, (1906), (1915-16), (1916), (1919), (1920), (1922), (1947)).

The work of Whitehead started in 1906 with the paper *On Mathematical Concepts of the Material World*, and one can also remember the relevant paper on *La Théorie Relationniste de L'Espace*, published in 1916. He then gave a complete solution to the problems of relationism in 1919-1920 by the books *An Enquiry on the Principles of Natural Knowledge* and *The Concept of Nature* (Whitehead, (1919), (1920), (1922)).

In 1903, only few years before Whitehead's solution, his scholar Bertrand Russell wrote, in the book entitled *The Principles of Mathematics,* that a relational theory of space and time should describe the principles of geometry in terms of sensible entities (Russell, (1903), § 395). Russell noted that indeed right lines and planes are not entities we encounter in sense experience, whereas, on the contrary, metrical (distance) relations are. Russell went on saying that indeed there is a very complicated method, invented by Leibniz and revised by Frischauf and Peano, by which only distance is fundamental, and the right line is defined from it, even if some of its properties can be introduced only by suitable axioms (Frischauf, (1872)¸ Couturat, (1901) 420; Peano, (1902-03)). The field of a given distance is the whole space, at variance with the field of the relation that gives rise to a right line which is only such right line itself. Such a relation generating the right line, hence, at variance with the former, makes an intrinsic distinction among space points, that is a distinction that a relational theory has to avoid. Pieri and others Peano's scholars have tried to formulate geometry starting from the fundamental concept of abstract motion, but they never create an entirely relational theory of geometry (Pieri, (1899)). This kind of approach to a relational theory of geometry did not start from actual physics and involved a change in the fundamental concepts of geometry: *metrical* geometry concepts replaced *descriptive* and *projective* geometry ones at the foundation level (Russell, (1897)). Whitehead's approach actually overcome this latter abstract (only mathematical) one.

However, after these works and Whitehead's theory, the relational question was almost completely hidden by the debate on general relativity, and specifically on the problem whether general relativity is actually a relational theory of space, time, and motion (Giannetto, (1987)). And it was also believed that this latter problem could be reduced to the technical problem of the embedding of the so-called "Mach's principle" within the framework of general relativity (Sciama, (1959); (1969), (1973); Wheeler (1964), (1988)). That is, by dealing with the misleading interpretation given by Einstein of Mach's idea of inertia (in Mach's perspective, it was due to the kinematical relation of each body to the remaining part of the universe, not to a dynamical (gravitational) effect) (Mach, (1883); Hoyle & Narlikar, (1974); Raine (1981)). Indeed, even if one accepts the pseudo-machian formulation of general relativity given by Sciama and others (Sciama, (1953, (1964); Sciama, Waylen & Gilman, (1969); Gilman, (1970); Lynden-Bell, (1967); Goenner, (1970); Reinhardt, (1972); Altschuler, (1985); Raine, (1975), (1981); Raine & Thomas, (1982); Raine & Heller, (1981)), a relational theory of space, time and motion is a more complex task than this reformulation of general relativity, a task which was realized for special relativity by Whitehead.

2. The Relational Theory of Space, Time and Motion: A Brief Account

Beyond Leibniz, Huygens and Mach, a relational conception of physics was at the roots of the theory of the actual creator of special relativity before

Einstein and of the four-dimensional space-time, that is (Jules) Henry Poincaré (Poincaré, (1880), (1898), (1900), (1902), (1904), (1905), (1906)). This priority was recognized by Hermann Minkowsky himself (Minkowsky, (1907/1916), and only by few physicists (Whittaker, (1953), 27-77; Tyapkin, (1972); Miller, (1973), (1981); Pais, (1982); Giannetto, (1995)). However, this kind of relational foundation - with the relevant exception of Eddington – (Eddington, (1920), (1923), (1928), (1938); Giannetto, (1994)) was almost completely lost in the formulation accepted by the scientific community as given by Einstein. However, one can say that neither Mach nor Poincaré themselves have developed such a deep, relational, understanding of the foundations of relativity as Whitehead.

It is well known that general relativity has turned upside down the hierarchy between kinematics (in some interpretation, dynamics) and geometry: the *kind* of *chrono*-geometry which enters in the construction of a physical theory is no longer given a priori, but it is defined by the kinematical, physical invariance group of transformations related to *kinematized* gravitodynamics (Barut, (1989); Finkelstein, (1969); Giannetto, (1991), (1993), (1994)). In this perspective, however, geometry has a foundation completely independent of physics at least at the *non-metrical* level, that is at the level of definition of point and straight line, even if *affine connection* is also defined by gravitation. Points and straight lines are mathematically constructed in a Platonist world of ideas: geometry is constructed on its own specific axioms regarding abstract concepts as points, lines, etc., and only after this stage physics could individuate by a very problematic choice only the kind of affine connection and metric, that is only the kind of *affine and metrical* geometry to be understood and used only as a physical application of already given mathematical structures. And even if one understands this determination of affine and metrical geometry by physics in a more radical way as the emergence of a *physical chrono-geometry* as opposed to *mathematical* geometry, only the affine connection and metrical structures, the superficial structures - one can say -, of geometry are physically determined, not the deep structures of geometry.

Only Eddington has had the idea to reduce *tout court* geometry to physics, in a relational perspective of chrono-geometry and of general relativity, but he has realized this reduction only *a posteriori*, by interpreting field equations of general relativity as an identity of metrical geometry functions (the $G_{\mu\nu}$ Einstein tensor) with physical functions (the $T_{\mu\nu}$ matter-energy tensor) (Eddington, (1923); Giannetto, (1994)). That is, such an identification happens only at a level of high-order (*non-fundamental*) geometrical and physical constructions.

Indeed, not only general relativity but also special relativity can be interpreted as involving the breakdown of the hierarchy between geometry and physics - apart from the Einstein's operational formulation (Einstein, (1905)), this point was recognized by Poincaré, Eddington and Whitehead. Here, the problem is the "embedding" of magnetic forces, and the definition of geometry is given by the kinematical invariance group of transformations related to *partially kinematized* electrodynamics (Giannetto, (1995)). Hence, already special relativity physics replaces a priori geometry with *chrono-geometry*, but also in this case it is only *metrical* geometry which is determined by physics.

In this perspective, one can understand how the question of relationism in relativity has been reduced to the technical satisfaction of the so-called Mach's principle: it is only a problem of the relation between two tensors, two non-fundamental variables. However, I would like to point out this conclusion: *Mach's principle is not sufficient for a relational theory of space, time and motion*. *Furthermore, in some sense, it is not even necessary*. Thus, we can have also a *relational* formulation of special relativity. On the other side, the general covariant formulation of special relativity (and indeed even of classical mechanics) satisfies some sort of "Mach's principle" (Havas, (1964), (1987); Logunov, (1990)).

I would like to show that one must come back to Whitehead's relational formulation of relativity (which - it must be repeated - is completely independent from his special-relativistic theory of gravitation as opposed to general relativity); then, also through the general covariant formulation of special relativity, one can automatically extend the relational formulation to general frameworks like general relativity too.

Whitehead, indeed, has solved the greatest question left by Leibniz: *relationism actually implies that every concept and every structure within a physical theory must be defined in terms of relations among physical "elements"; no mathematical or logical concept or structure can be given independently from physical relations*. *Every other option leads to meta-physics. There is no conventionality of metric* (Giannetto, (1993)). The fundamental concepts of physics like space and time cannot have any mathematically or logically given *a priori* structure.

In Whitehead's formulation of special relativity, physics not only defines the *metrical* geometry, but it also defines *non-metrical, affine, descriptive or projective* geometry, that is geometry *tout court* from its "foundations". Physics defines geometry not only *a posteriori*, at the level of high-order constructions as field equations like in Eddington's interpretation of general relativity, but physics defines points, lines, planes and so on, in terms of fundamental physical processes, that is not in terms of relations among *bodies* or high-level tensors (*matter*), but in terms of relations among *event-particles* (Whitehead, 1922). From this point of view, only Whitehead's relational chrono-geometry is an actual *physical* geometry, free from any logico-mathematical (Platonist or Kantian, any way idealistic) presuppositions.

Let us consider, first of all, relationism in respect to the fundamental concepts of geometry. Already in 1906 paper, Whitehead was pointing out that the simplicity of spatial points was in opposition to the relational theory of space: this requires points to be non-fundamental, complex entities (Whitehead, (1920). The statement that the event-particle which one can coordinatizes by four quantities (p_1 , p_2 , p_3 , p_4) occupies or happens in the point (p_1 , p_2 , p_3) means only that the event-particle is only one of the series of event-particles which is the point. That is, a spatial point is only a series, a set of physical event-particles which have in common the first three quantities. Hence, a theory of space is not a theory of relations of objects, but of relations of events. Whitehead explained that in the orthodox theory events are described by means of objects which occupy a dominant position, and so events are considered as a mere play of relations among objects. In this way space theory becomes a theory of relations among objects instead of relations among events. The consequence is that, for objects are not related to the becoming of events, space as relations among objects is considered as unconnected to time. But there cannot be space without time, or time without space, or space and time without event becoming (Whitehead, 1922)). Thus, at variance with the major part of interpretations of relativity which speak about the spatialization of time, Whitehead obtained a complete *temporalization of space*, so overcoming all the philosophical criticism about that seeming feature of relativity (see also Capek, (1961)). An idea of temporalization of space was already present in Leibniz (Poser, (1993)).

Whitehead wrote in *The Principle of Relativity with applications to Physical Science*:

> [...] Nature is stratified by time. In fact, passage in time is of essence of Nature, and a body is merely the coherence of adjectives qualifying the same route through the four-dimensional space-time of events. But as the result of modern observations we have to admit that there are an indefinite number of such modes of time stratification. However, this admission at once yields an explanation of the meaning of the instantaneous spatial extension of nature. For it explains this extension as merely the exhibition of the different ways in which simultaneous occurrences function in regard to other time-systems. I mean that occurrences which are simultaneous for one time-system appear as spread out in three dimensions because they function diversely for other time-systems. The extended space of one time-system is merely the expression of properties of other time-

systems. According to this doctrine, a moment of time is nothing else than an instantaneous spread of nature. Thus, let t_1, t_2, t_3 be three moments of time according to one time-system, and let T_1, T_2, T_3 be three moments of time according to another time-system. The intersection of pairs of moments in diverse time-systems are planes in each instantaneous three-dimensional space... (Whitehead, (1922): 54-55)

In a more synthetic way, he had written in the introduction:

Position in space is merely the expression of diversity of relations to alternative time-systems. Order in space is merely the reflection into the space of one time-system of the time-orders of alternative time-systems. A plane in space expresses the quality of the locus of intersection of a moment of the time-system in question (call it 'time-system *A*') with a moment of another time-system (time-system *B*). The parallelism of planes in the space of time-system *A* means that these planes result from the intersections of moments *A* with moments of one other time-system *B*. A straight line in the space of time-system *A* perpendicular to the planes due to time-system *B* is the track in the space of time-system *A* of a body at rest in the space of time-system *B*. Thus, the uniform Euclidean geometry of spaces, planeness, parallelism, and perpendicularity are merely expressive of the relations to each other of alternative time-systems. The tracks which are the permanent points of the same time-system are also reckoned as parallels. Congruence - and thence, spatial measurements - is defined in terms of the properties of parallelograms and the symmetry of perpendicularity. Accordingly, position, planes, straight lines, parallelism, perpendicularity, and congruence are expressive of the mutual relations of alternative time-systems (Whitehead, (1922): 8-9).

Let us consider now properly kinematics. Motion is another relation of events, that is a series of events (p_1, p_2, p_3, p_4) linked to an object, conceived as placed in them, which is defined by its relation with the *remaining part* of the universe. If one considers another time-system (reference frame), the same motion will appear as a relation of *other* events (q_1, q_2, q_3, q_4), which in general are associated to other different objects. Hence, even if the motion of one object is relative to the particular considered time-system, such a motion cannot be reduced to an overall rest in any other time-system: that is, it will transform itself into the motion of the *remaining part* of the universe (Whitehead, (1920)). Thus, one must say that Whitehead rejected only the Einsteinian content of the so-called Mach's principle, not its kinematical actual (Machian) meaning. Indeed, Whitehead kinematized the concept of physical field of an object: it is nothing else than the collection of modifications of event series related to that object: it is a kinematical relation among events and it does not involve any contact or at-a-distance action (his theory of gravitation was not conceived as an action at-a-distance theory as often stated).

Therefore, motion is a relation and is relative to a time-system, but it has a real counterpart. This furnishes us with a Leibnizian interpretation of relativity: the subject of motion is not an invariant, but overall motion is.

3. Relativity as a Physical Hermeneutics

One can understand better Whitehead's work by schematizing and comparing in the following way the different kind of constructions of the physical theories:

<u>Classical Mechanics</u>	<u>Poincaré's Special Relativity</u>
	- experiments
- epistemology and ontology	- epistemology and ontology
- logic	- logic
- set theory	- set theory
- topology	- topology
- non-metrical geometry	- non-metrical geometry
- metrical geometry of bodies	- electrodynamics of fields
- kinematics	- kinematics
- dynamics	- metrical chrono-geometry
- verification experiments	

<u>Whitehead's Special Relativity</u>	<u>Einstein's General Relativity</u>
- lifeworld experience/ experiments	
- epistemology and ontology of interrelated events	- epistemology and ontology of bodies
- (electro-)kinematics of events	
- logic of events	- logic
- set theory of events	- set theory
- topology of events	- topology electrodynamics
- non-metrical chrono-geometry	- non-metrical geometry kinematics
	- gravito-dynamics
- event metrical chrono-geometry	- body metrical chrono-geometry
- (gravito-)kinematics	pseudo-Riemannian
	- verification experiments

These four schemas represent very different hierarchies of steps from the top to the bottom in the construction of physical theories (Finkelstein & Rodriguez, (1983); Giannetto, (1991)). Except Whitehead's case, in the other ones the steps from epistemology to non-metrical geometry (and for classical theories indeed up to kinematics) are almost completely unquestioned presuppositions to a physical theory, that is *meta-physical* presuppositions.

Even if Poincaré (at variance with Einstein and Minkowski) had discussed (giving many contributes) practically all the problems related to such steps, he left these levels as untouched by relativistic physics.

Whitehead is the only one to derive all these levels (not only non-metrical geometry) from the consideration of physical processes (events), trying to overcome the foundationalist paradigm of an epistemological or ontological (that is, subjectivistically or objectivistically meta-physical) ultimate ground for knowledge and physics (Rorty, (1979); Giannetto, (1991)). His starting point, as I have emphasized, is also experience (Giannetto, (2010)) with the reconsideration of the full experience and of experiments, but without any transcendental foundation; *relationism is not a mere relativism or a particular epistemological option but a sort of a physical hermeneutics* (Giannetto, (1991)). In fact, Whitehead has given us the deepest conception of relativity: in his approach, the principle of relativity is first of all, actually, an ontological principle of universal *relatedness* of Nature; the *indeterminateness* of the subject of motion is based, with all its epistemological implications, upon this ontological relatedness. Nature is not a simple aggregate of independent and separate entities: the traditional mechanistic view has represented Nature as an accidental system of contingent separate entities; however, *relativity as relationism shows that events are non-separable within the world as a whole* (Whitehead, (1922); Giannetto, (1995)).

The appearance of an *entirely physical (theoretical) practice* represents an epochal change, *an epochal departure from Western meta-physics* (Giannetto, (2010)). However, as it can be seen from his construction of special relativity, gravito-kinematics was left by Whitehead into one of the bottom levels, at variance with general relativity: Whitehead recognized that a choice like the one operated in general relativity construction would lead to an actual hidden breakdown of the metrical geometry structure (Whitehead, (1922); Giannetto, (1994)). Thus, Whitehead's choice in this respect was not good from a radical relationist point of view, just indipendently of the validity of general relativity. It should not be so difficult to elaborate, on one side, a complete relationist theory of gravitational processes too in an actual Whitehedian form, and, on the other side, in any case it is easy to give a relational construction to general relativity, by considering, just on the same level of matter event-particles, gravitational event-particles too.

It is so clear that Whitehead's formulation of special relativity is not equivalent to Poincare's or the other traditional version of the theory at least from an epistemological point of view; but, indeed, also from a mathematical point of view, the structure of the theory is different until up to the metrical geometry level. Therefore, differing at an epistemic and mathematical level, and furthermore at the semantic level (for example, the idea of temporalization of space), Whitehead's special relativity seems to be a new, different physical theory more than a mere reformulation of Poincaré's or Einstein-Minkowski's special relativity (Ushenko, (1949)). However, the observational consequences to be related to the metrical geometry structure are identical and Whitehead's special relativity indeed gives us *a completely relational physical theory* in which we no longer appear as having to include the world but we are included in the world.

4. The Relationality of Motion and the Relatedness of Nature

The principle of relativity, according to Whitehead, is an ontologicalprinciple, not only an epistemological one (Whitehead, 1922): the impossibility of knowing the subject of motion is the consequence of the *universal relatedness* of *Nature*, an ontological principle of inter-relationship of every material body with all other material bodies, which so holistically constitute Nature. This constitutive interrelation of all material things explains why our knowledge has limitations in defining individual properties of bodies. Relations between bodies are not "ideal" relations (as in Leibniz) introduced by the human intellect to order them, but they are real: the fields of forces exist even in the absence of material bodies.

Whitehead gives a new interpretation of the "principle of (special) relativity of motion", which tells us that, in the absence of a certainly fixed reference, at rest, it is impossible to know observationally or theoretically which body, between two bodies in reciprocal motion is at rest and which one is not at rest (in rectilinear and uniform motion), or what is the "subject of the motion".

If everything was at rest, everything would appear at rest; if a body for one reference system appears at rest and for another reference appears in motion, then either the body or the reference must be in motion: the relativity of motion implies thatat least one *motion exists*. If the Earth is considered at rest, the Sun is moving: motion is projected on another body. In every reference frame system, there is something in motion, something changes. Even if for a body we cannot know whether it is in motion or at rest, we know that there is motion in Nature: motion as a relation is absolute, is invariant. Rest is only a relative rest, that is the situation in which two bodies have the same motion.

A transformation of reference frame transforms a motion of a body in a motion of another body (we have not the same events), but it conserves a certain temporal succession structure of events which we call motion.

The relativity of motion would not occur if there were only static reference frame systems (at rest): it is the consequence of the possibility/need to consider reference frame systems in motion.

Things do not change if we consider non-inertial motions and non-inertial reference frame systems, for which we can state a principle of general relativity of motion.

The principle of general relativity of motion tells us that it is impossible to know observationally or theoretically which body,between two bodies in reciprocal motion, is at rest and which is in motion (even accelerated in any way), or what is the "subject of motion": then or the body or the reference frame system must be in motion: the general relativity of motion implies that at least one *motion* accelerated in any way *exists*. General arbitrary transformations of reference frame systems can alter the rest or the kind of motion of a body, but they transform an arbitrary motion of a body in the same kind of motion of another body: in different reference frame systems we have not the same events (concerning the same bodies), but a temporal succession structure of events (concerning different bodies) which identifies motion is preserved in the transformations (Whitehead, (1920)).

In Newtonian modern physics, a body or a reference frame system is in an accelerated motion only if a force field acts on it, a field that accelerates it: experimentally, on every body or material reference frame system acts a field of gravitational forces, because gravitation is universal. Strictly speaking all bodies – unless the gravitational field is artificially cancelled – and so all reference frame systems move in accelerated motion because they interact with all other bodies in the universe through gravitational force fields. An accelerated reference frame system, that can relativistically modify motion making it a relative thing, can exist only because there are interactions that realize a universal relatedness of Nature. There is no body or reference frame in absolute rest and a general relativity of motion is given.

The principle of general relativity of motion is the consequence of an unavoidable "solidarity of the universe", realized through a 'universal relatedness of Nature', i.e. a 'universal relationality of Nature', a universal field of (cor-)relations. There are no isolated and separable bodies: Nature is a totality of non-separable parts.

We can have general arbitrary transformations of reference frame systems which can alter the rest or the kind of motion of a body, and which transform an arbitrary motion of a body in the same kind of motion of another body and preserve a temporal succession structure of events, concerning different bodies, only because of the universal relatedness of Nature. A change in a part of Nature must imply a change in another part (Whitehead, 1920).

The principle of relativity is a principle that establishes our ignorance, an epistemological principle that concerns first of all a limit of our knowledge: in general, we cannot attribute to a single body motion as its individual property, but we can only establish it as a relationship between two bodies. We can know only in some special cases, concerning us as moving bodies, which body is moving, but motion is always a relation of a body to other bodies: for a unique existing body, we could not distinguish motion and rest. Motion is a property of Nature as a whole.

This fact, that motion for us is attributable to a body not as an individual property but only as a relationship with another body or with other bodies or relative to a certain chosen point of view (to a "frame system of reference"), leads Whitehead to conclude that in general we cannot abstract a material body from the existence of the other material bodies with which it is by nature related, i.e. that the universe is not made up of separable material bodies, but rather by bodies that cannot be separated from each other. *Being in relation to other bodies constitutes the essence of a body* and therefore one must consider the universe as an inter-related whole.

5. No substantial material bodies but events. Nature as a whole temporal process

However, there's more. The very concept of an individual material body separable from others loses its consistency and can no longer be the basis on which we can constitute the idea of Nature.

If we can affirm that there is a certain relationship of motion between two bodies that can never be completely eliminated, because, even when, from a certain point of view, from a certain frame system of reference, a body is at rest, we must attribute motion to another body – that is, either it is in motion one or the other – what is truly real (invariant for all the reference frame systems) is not the individual body with its supposed properties of motion or stillness that we cannot ascertain, but rather *motion as a relational (collective) property of Nature*. We cannot conceive a body without definite properties, it would be an abstraction. Motion is only a series of events: it is not something identifying or not a body.

Nature then is not made up of stable separable individual material bodies, but Nature is motion as a relation of the parts as events: change, process.

We must include individual material bodies only as *relative parts* of a process, of a change-motion, which, as such, can never be described only in spatial terms, but always implies also a *space-time dimension*: *a temporal series of events.*

It then explains why in the theory of relativity we must move to a physical description *in a four-dimensional space-time*: because the *Nature* to be described is not made of individual separable, stable in some spatial position, material bodies, but rather *is made of motions, changes, processes, events.*

There are *no more things-in-themselves-substances but only (fields of) events*.

Different relations of motion between different parts of Nature imply different temporal relations. Nature is a set of different processes-motions, *a set of different temporalities*. Nature is not a timeless reality, as in contemporary quadridimensionalist metaphysics.

We understand that *space must also be rethought in terms of time* and we can also understand it in our experience if we do not make abstractions. The weft of space is woven by the vertical warp of the times.

A point in timeless space is not a fundamental entity, but it is *thehistorical-temporal set of events*, of the processes that happened there: $P = [e_1, e_2, e_3, e_4, ...]$

This kind of conception is in close agreement also with our actual experience of space and time. For example: what is a city like Messina? Is it just a spatial place that we can know by means of geography? Is it just where we are now? That kind of definition would be reductive. Isn't it also the place where my parents, ancestors, or other people lived in the past? Isn't it also the republic in the seventeenth century subjugated by the ferocious Spaniards? Isn't it also the place where the Turks killed and ruled? Isn't it also the place where the Greeks from Messana in Greece partly moved? Isn't it also the place where, after us, our children or other people will live? We understand that spatial geography is not enough to define Messina, but rather there is a need to add a historical-temporal dimension to define it. A place like Messina, as well as a point in space, is not a fundamental entity, but it is the historical-temporal set of events, of the processes that took place there.

What is a house? Is it just a place we live in now? No, to say that would be an abstraction. The house where I live now is also the house that belonged to my parents and will be my heir's when I shall die. The house is a time series of events, of processes, not a material building in urban space.

What is physically a sofa? Is it just the present sofa where I'm sitting right now? No, it's also the place where my parents sat in the past, and where maybe others will sit after me in the future, if it's not thrown away.

It is well known that general relativity has *turned upside down the hierarchy between kinematics (in some interpretation, dynamics) and geometry*: the *kind* of *geometry* which enters in the construction of a physical theory is *no longer given a priori*, but it is defined by the kinematical, physical invariance group of transformations related to *kinematized* gravitodynamics (Barut, 1989).

In this perspective, however, geometry has a foundation completely independent of physics at least at the *non-metrical* level, that is at the *affine* or *projective* geometrical level. It is mathematically constructed in a Platonist world of ideas, on its own specific axioms regarding abstract concepts as points, lines, etc., and only after this stage physicscould individuate by a very problematic choice only the kind of metric, that is only the kind of *metrical* geometry to be understood and used only as a physical application of already given mathematical structures.

And even if one understands this determination of metrical geometry by physics in a more radical way as the emergence of a *physical chrono-geometry* as opposed to *mathematical* geometry, it is only the metrical structure of geometry that is physically determined.

Thus, what is *a material body* in general? It *is a time series of events*, of processes. Nature is the process of all the interrelated processes. *The visible space for us, given the finite speed of light, is not only what happens in our present, but the set of different pasts of all the other processual temporalities of all the other parts of the universe: visible space is the unfolding of different times.*

6. Whitehead’s Interpretation Against Some Recent Eternalist Philosophies of Special Relativity

Whitehead’s interpretation of relativity can be used to refute some recent philosophies: eternalism, fourdimensionalism, perdurantism, endurantism, exdurantism (Calosi, (2015). According to Whitehead, reality is, as the medieval philosophers of motion said, a *res successiva,* never a timeless object.

Four-dimensionalist philosophy proposes to consider all the temporal phases-parts of processes as *coexisting not* (?) *simultaneously* and to consider as real four-dimensional objects extended in time as well as in space, reducing time to spatial extension and thinking of such objects as persistent in time. Indeed, the notion of coexistence implies simultaneity. It is a matter of considering temporal succession as illusory and time as unfolded as simultaneously: past, present and future would always coexist simultaneously in a vision linked to an eternalism opposed to presentism that considers only the present real. You would have a block *time* or a block *universe* as an immutable four-dimensional block. The prospect of considering past, present and future coexisting (indeed, simultaneously) is not justifiable on the basis of the impossibility of establishing in relativity a temporal order, invariant for all reference systems, for non-causally connectable events (events linked by a *space-like space-time* interval), nor on the basis of the sole authority of Albert Einstein who has been appealed, by Federigo Enriques and Karl Raymund Popper, as the "new Parmenides" for his refusal to consider temporal succession real. The impossibility of establishing an invariant temporal order for certain events (*spacelike*) and not for all implies only the incommensurability of different temporal sequences of events at different points in space, and never a real coexistence (indeed, simultaneity) of all events: such different temporal sequences of different physical systems can however be compared and ordered temporally within the broader order of the temporal sequence of a system that it comprises as parts and includes all the events of the spacelike timelines in a *time-like* interval relation. The order of time is local, as in the case where, while for Galileo the life of Lucretius belonged to the past, for Epicurus it was part of the future. The order of time is also local because time is flowing. However, if we consider the history of mankind as a timeline, today we can include in its past both Galileo, Epicurus, and Lucretius.

At the basis of relativity is the loss of meaning of the possibility of establishing simultaneity at a distance in space, with the consequence that a distance in space must instead be interpreted in terms of a temporal sequence of events. Eternalism is thus a logical and physical fallacy and constitutes a total misunderstanding of the processual-temporal character of four-dimensional space-time, as explained by Whitehead.

Following the four-dimensionalist philosophy, the Lorentz contraction is not real but is only a three-dimensional projection of reality: in the three dimensions there is the contraction of lengths with the dilation of time that can be real due to the magnetic field or it can be simulated by the change of the reference system; but the instantaneous section is an abstraction always because the instant does not exist and there is no simultaneity at a distance and an instantaneous space. Contraction is always related to a dilation of a time interval and therefore is not related to an instantaneous section: four-dimensionalist interpretation is thus mistaken. That the contraction and dilation then change or disappear in a particular other reference is obvious: the four-dimensional space-time volume is the same, but it indicates our ignorance about true time and true space (space is time and varies according to the rhythm of the time of the process and the reference, space-time is not a 4-dimensional space but Whitehead showed it as a temporally characterized space; the volume therefore depends on the rhythm of time with which it is measured, but if we multiply it by time, the gamma factor is elided in $\Delta\mathbf{r}\ \Delta t$ and we get the volume for the proper time of the process. If space is full of objects is given, but empty space does not exist: the empty space that exists, for example, between two celestial bodies, is *only that which can be traveled or that is actually traveled*. It is a different thing if one travels through it with a motion at a certain speed or with another motion at another speed. Time defines space: this definition is such that the space-time interval or the space-time product is invariant, because it is the one defined by light in a univocal way as the distance between two events or as the "evolutionary volume" of a system of certain spatial dimensions in time). Space-time as an interval indicates proper time, which we do not know what it is. What is invariant is the motion of light that corresponds to a space-time volume for which the magnetic field may or may not exist.

Four-dimensionalist eternalism has been declined in two versions. *Perdurantism*, which imagines reality as given by the mereological set of all the temporal parts that make up a single four-dimensional object that endures, referred to as a space-time *worm,* like the enduring set of various rings corresponding to the temporal parts; and *exdurantism* (*exdurantism*), according to which persistent objects are the individual temporal parts (time-slices, *instantaneous temporal slices* derived from cuts in *space-time*) that constitute individual *stages* that bind together in a gen-identity relationship.

Endurantism, on the other hand, is the philosophical perspective according to which persistent objects are three-dimensional material objects that are completely present in every moment of their existence: this perspective is linked to an A-theory of time (in which time is thought of as a continuous transformation of events from future to present to past) by John McTaggart (1908); while perdurantism is linked to a *B-theory* of time (in which events are not thought of in their flow but in *tenseless* relationships, i.e. without the temporal specifications of past, present and future, and therefore static-spatial of "before" and "after" that remain stable).

Both endurantism, perdurantism and ex-durantism postulate the persistence of objects, respectively three-dimensional or four-dimensional, whereas, according to Whitehead, the theory of relativity implies a physical reality given by a temporal processuality of non-persistent events.

The 4-dimensional space-time was introduced in 1905 by Henri Poincaré within the new relativistic electromagnetic dynamics: its introduction was necessary in the perspective in which it was demonstrated the possibility of understanding the phenomena related to material bodies in terms of phenomena of the electromagnetic field; the electromagnetic field, consisting of electromagnetic waves, is a form of motion. The movement of electromagnetic waves cannot be described only in a static spatial geometric framework but also requires time, a temporal dimension. While material bodies can be at rest at a certain moment and a three-dimensional geometry that allows them to be placed in a certain position at a certain point in space can be enough to describe their state, electromagnetic waves, being a form of wave motion, necessarily also require the temporal dimension to be described. and, therefore, the replacement of a three-dimensional geometry with a new four-dimensional "chrono-geometry". The temporal dimension is thus recognized as constitutive of physical reality. Physical reality is no longer given by material bodies that can also be considered at rest, but by temporal processes (electromagnetic waves): four-dimensional space-time is the description of temporal processes.

7. Conclusions

Whitehead's Special Relativity is so hierarchically structured: lifeworld experience (experiments too); epistemology and ontology; *Relativistic logic of events; Relativistic set theory of events; Relativistic number theory of events; Relativistic topology of events; Relativistic non-metrical chrono-geometry; Relativistic metrical chrono-geometry.*

Thus, Whitehead realized a relational reformulation of logic (against the metaphysics of subject-predicate logic related to the metaphysics of substances), a relational reformulation of mathematics (set theory, arithmetic, algebra, topology, non-metrical and metrical geometry), a physical reformulation of logic and mathematics.

The physics of relativity makes us understand the temporal and processual reality of things and Whitehead's philosophy.

Whitehead's special-relativistic theory of gravitation can be understood not as an alternative to Einstein's general relativity theory, but in terms of a special-relativistic limit of the general relativity theory of gravitation. General relativity has been formulated as having two limits to which it reduces itself to previous theories: locally in space-time general-relativistic dynamics reduces itself to special-relativistic dynamics; and furtherly, in the limit of weak fields, general relativistic theory of gravitation reduces itself to the Newtonian theory of gravitation. Thus, the limit of general relativity is schizophrenic: Whitehead' special-relativistic theory of gravitation filled a structural gap and made possible to consider special relativity as the unique limit of general relativity theory.

Following Whitehead, for the relationality of Nature, each part is involved in everything: one part is the set of all relations with the rest of the universe (*togetherness*): it is the relationship with all the other parts, with the otherness that constitutes every part of the universe. *Nature is an inter-related totality*: it is therefore not like a machine, but constitutes a living organism. Every part of Nature is sensitive to the others, every part is alive in different degrees. A *new non-mechanist image of Nature*.

Process and Reality (Whitehead, 1929) can be understood in terms of Whitehead's interpretation of relativity.

The relational ontology of Nature implies a *cosmic relational ethics*, respectful of all other parts of the universe, of every living part. One new relational image of God as a love that grows with always new relationships of the creative process of the universe.

References

Alexander, H. G. (1984), *The Leibniz-Clarke Correspondence*, New York, Barnes & Noble.

Altshuler, B. L. (1985), "Mach's Principle. Part 1. Initial State of the Universe", in *International Journal of Theoretical Physics*, v. 24 pp. 99-118.

Barut, A. O. (1989), *Geometry and Physics. Non-Newtonian Forms of Dynamics*, Napoli, Bibliopolis.

Calosi, C. (2015), *Relatività Speciale e Metafisica Analitica*, in Pecere, P. (ed.) (2015), *Il Libro della Natura,* I-II, Roma, Carocci, vol. II, pp.85-113.

Capek, M. (1961), *The Philosophical Impact of Contemporary Physics,* Princeton (N.J.), van Nostrand.

Couturat, L. (1901), *La Logique de Leibniz*, Paris, Alcan.

Earman, J. (1989), *World Enough and Space-Time. Absolute versus Relational Theories of Space and Time*, Cambridge, Mass., The MIT Press.

Eddington, A. S. (1920), *Space, Time and Gravitation. An Outline of the General Relativity Theory*, Cambridge University Press, Cambridge.

Eddington, A. S. (1923), *The Mathematical Theory of Relativity*, Cambridge, Cambridge University Press.

Eddington, A. S. (1928), *The Nature of the Physical World*, Cambridge, Cambridge University Press.

Eddington, A. S. (1938), *The Philosophy of Physical Science*, Cambridge, Cambridge University Press.

Einstein, A. (1905), "Zur Elektrodynamik bewegter Körper", in *Annalen der Physik*, s. 4, v. 17, pp. 891-921.

Finkelstein, D. (1969), "Matter, Space and Logic", in *Boston Studies in the Philosophy of Science*, v. 5, p. 199.

Finkelstein, D. & Rodriguez, E. (1983), *Quantum Simplicial Topology*, GIT preprint.

Frischauf, J. (1872), *Absolute Geometrie nach Johann Bolyai*, Leipzig, Teubner.

Giannetto, E. (1987), "Il crollo del concetto di spazio-tempo negli sviluppi della fisica quantistica: l'impossibilità di una ricostruzione razionale nomologica del mondo", in *Aspetti epistemologici dello spazio e del tempo*, ed. by Boniolo, G., Roma, Borla, pp. 169-224.

Giannetto, E. (1989), "Fisica Quantistica e Verità Logica", in *Epistemologia* XII, pp. 261-276.

Giannetto, E. (1991), "On Truth: A Physical Inquiry", in *Atti del Congresso Internazionale "Nuovi Problemi della Logica e della Filosofia della Scienza"*, ed. by Cellucci, C. & Dalla Chiara, M., Clueb, Bologna, v. I, pp. 221-228.

Giannetto, E. (1991), "L'epistemologia quantistica come metafora antifondazionistica", in *Immagini Linguaggi Concetti*, ed. by Petruccioli, S., Roma, Theoria, pp. 303-322.

Giannetto, E. (1993), "The Epistemological and Physical Importance of Gödel's Theorems", in *First International Symposium on Gödel's Theorems*, ed. by Z. W. Wolkowski, Singapore, World Scientific, pp. 136-147.

Giannetto, E. (1994), "Note sull'interpretazione della relatività generale di A. S. Eddington", in *Atti dell'XI Congresso Nazionale di Storia della Fisica*, ed. by Bevilacqua, F., Pavia, Goliardica Pavese, pp. 121-135.

Giannetto, E. (1994), "Mach's Principle and Whitehead's Relational Formulation of Special Relativity", in *Proceedings of the Conference on the Physical Interpretations of Relativity Theory III*; London, Unicam, pp. 126-146.

Giannetto, E. (1995), "Henri Poincaré and the rise of special relativity", in *Hadronic Journal Supplement* 10, pp. 365-433.

Giannetto, E. (1995), "Some Remarks on Non-Separability", in *The Foundations of Quantum Mechanics*, Garola, C. & Rossi, A. (eds.), Dordrecht, Kluwer, pp. 315-324.

Giannetto, E. (2005), *Saggi di storie del pensiero scientifico*, Bergamo, Bergamo University Press, Sestante.

Giannetto, E. (2010), *Un fisico delle origini. Heidegger, la scienza e la Natura*, Roma, Donzelli.

Gilman, R. C. (1970), "Machian Theory of Inertia and Gravitation", in *Physical Review* D, v. 2, p. 1400.

Goenner, H. (1970), "Mach's Principle and Einstein's Theory of Gravitation", in *Boston Studies in the Philosophy of Science*, v. 6, Dordrecht, Reidel.

Grünbaum, A. (1973), *Philosophical Problems of Space and Time*, Dordrecht, Reidel.

Havas, P. (1964), "Four-Dimensional Formulations of Newtonian Mechanics and Their Relation to the Special and the General Theory of Relativity", in *Reviews of Modern Physics*, v. 36 pp. 938-965.

Havas, P. (1987), "Simultaneity, conventionalism, general covariance, and the special theory of relativity", in *General Relativity & Gravitation*, v. 19, pp. 435-453.

Hoyle, F. & Narlikar, J. (1974), *Action at a distance in physics and cosmology*, San Francisco, Freeman.

Huyghens, C. (1905), *Oeuvres complètes de Christian Huyghens*, Aja v. X (correspondence 1691-1695), p. 609.

Jammer, M. (1954), *The History of Theories of Space in Physics*, Cambridge, Mass., Harvard University Press.

Jammer, M. (1957), *Concepts of force. A study in the foundations of dynamics*, Cambridge, Mass., Harvard University Press.

Kitchener, R. F. (ed.) (1988), *The World View of Contemporary Physics*, Albany, State University of New York Press.

Korteweg, D. J. & Schouten, J. A. (1920), "Huygens", in *Jahresbericht der Deutschen Mathematiker-Vereinigung*, v. 29, p. 136.

Leibniz, G. W. (1849), *Leibnizens mathematische Schriften*, ed. by Gerhardt, C. I., Halle, Asher.

Leibniz, G. W. (1875), *Die philosophischen Schriften von G. Leibniz*, ed. by Gerhardt, C. I., Weidemann, Berlin & Hildesheim, Olms, 1960.

Logunov, A. (1990), *Lectures in Relativity and Gravitation. A Modern Look*, Moscow, Nauka, & Oxford, Pergamon Press.

Lowe, V. (1941), "The Development of Whitehead's Philosophy", in P. A.

Schilpp (ed.) (1941), *The Philosophy of A. N. Whitehead*, Evanston and Chicago, The Library of Living Philosophers, Northwestern University, pp. 15-124.

Lynden-Bell, D. (1967), "On the Origins of Space-Time and Inertia", in *Monthly Notices of the Royal Astronomical Society*, v. 135, pp. 413-428.

Mach, E. (1883), *Die Mechanik in ihrer Entwickelung historisch-kritisch dargestellt*, Leipzig, Brockhaus.

Miller, A. I. (1973), "A Study of Henry Poincaré's *Sur la dynamique de l'électron*", in *Archives for the History of Exact Sciences*, v. 10 pp. 207-328.

Miller, A. I. (1981), *Albert Einstein's Special Theory of Relativity: Emergence (1905) and Early Interpretation (1905-1911)*, Reading Mass., Addison-Wesley.

Minkowski, H. (1907), "Das Relativitätsprinzip", Lecture delivered on 5 November 1907, in *Annalen der Physik*, IV Folge, v. 47 (1915), pp. 927-938.

Northrop, F. S. C. (1941), "Whitehead's Philosophy of Science", in P. A. Schilpp (ed.) (1941), *The Philosophy of A. N. Whitehead*, Evanston and Chicago, The Library of Living Philosophers, Northwestern University, pp. 165-207.

North, J. D. (1965), *The Measure of the Universe. A History of Modern Cosmology*, Oxford, Oxford University Press.

Ockham, Guillelmi de (1984-5), *Opera Philosophica*, voll. IV-V-VI, New York, St. Bonaventure University.

Pais, A. (1982), *'Subtle is the Lord...'. The Science and the Life of Albert Einstein*, Oxford, Oxford University Press.

Peano, G. (1902-03), "La geometria basata sulle idee di punto e distanza", in *Atti della Reale Accademia delle Scienze di Torino*, v. XXXVIII, pp. 6-10.

Pieri, M. (1899), "Della geometria elementare come sistema ipotetico-deduttivo", in *Memorie della Reale Accademia delle Scienze di Torino*, v. XLIX, p. 176.

Poincaré, H. (1880), "Note sur les principes de la mécanique dans Descartes et dans Leibnitz", in Leibnitz, G. W., *La Monadologie*, ed. by Boutroux, E., and with a note by Poincaré, H., Paris, Delagrave, pp. 225-231.

Poincaré, H. (1898), "La mesure de temps", in *Revue de Métaphysique et Morale*, v. 6 pp. 1-13.

Poincaré, H. (1900), "La théorie de Lorentz et le principe de réaction", in *Arch. Néerl.*, v. 5, pp. 252-278; reprinted in *Recueil de travaux offerts par les auteurs à H. A. Lorentz*, The Hague, Nijhoff, 1900.

Poincaré, H. (1902), *La Science et l'Hypothèse*, Paris, Flammarion, 1902, 2nd ed. 1907.
Poincaré, H. (1904), "L'état actuel et l'avenir de la Physique mathématique", in *Bulletin des Sciences Mathematiques*, v. 28, pp. 302-324.
Poincaré, H. (1905), "The Principles of Mathematical Physics", in *The Monist*, v. 15, p. 1.
Poincaré, H. (5 June1905), "Sur la dynamique de l'électron", in *Comptes Rendus de l'Académie des Sciences*, v. 140, pp. 1504-1508.
Poincaré, H. (23 July 1905), "Sur la dynamique de l'électron", in *Rendiconti del Circolo Matematico di Palermo*, v. 21 (1906), pp. 129-175.
Poser, H. (1993), *La teoria leibniziana della relatività di spazio e tempo*, aut-aut, v. 254-255, pp. 33-48.
Raine, D. J. (1975), "Mach's Principle in General Relativity", in *Monthly Notices of the Royal Astronomical Society*, v. 17,1 pp. 507-528.
Raine, D. J. (1981), "Mach's principle and space-time structure", in *Report on Progress in Physics*, v. 44, pp. 1151-1195.
Raine, D. J. & Heller, M. (1981), *The Science of Space-Time*, Tuscon, Pachart Publishing House.
Raine, D. J. & Thomas, E. G. (1982), "Mach's Principle and the Microwave Background", in *Astrophysics Letters*, v. 23, pp. 37-45.
Reichenbach, H. (1924), "Die Bewegungslehre bei Newton, Leibniz und Huyghens", in *Kantstudien*, v. 29, pp. 416-438.
Reinhardt, M. (1972), "Mach's Principle. A critical Review", in *Zeitschrift für Naturforschung*, v. 28 A, pp. 529-537.
Russell, B. (1897), *An Essay on the Foundations of Geometry*, Cambridge, Cambridge University Press.
Russell, B. (1900), *A Critical Exposition of the Philosophy of Leibniz*, London, G. Allen & Unwin Ltd.
Russell, B. (1903), *The Principles of Mathematics*, Cambridge, Cambridge University Press.
Russell, B. (1914), *Our Knowledge of the External World*, La Salle, Open Court.
Russell, B. (1927), *The Analysis of Matter*, London, G. Allen & Unwin Ltd.
Schild, A. (1956), "On gravitational theories of Whitehead's type", in *Proceedings of the Royal Society of London*, v. 235, pp. 202-209.
Schilpp, P. A. (ed.) (1941), *The Philosophy of A. N. Whitehead*, Evanston and Chicago, The Library of Living Philosophers, Northwestern University.
Sciama, D. W. (1953), "On the origin of inertia", in *Monthly Notices of the Royal Astronomical Society*, v. 113, p. 34.
Sciama, D. W. (1959), *The Unity of the Universe,* London, Faber and Faber.

Sciama, D. W. (1964), "The Physical Structure of General Relativity", in *Reviews of Modern Physics*, v. 36, pp. 463-469.
Sciama, D. W. (1969), *The Physical Foundations of General Relativity*, London, Heinemann.
Sciama, D. W. (1973), *Cosmology now*, ed. by Laurie, J., London, BBC Publications.
Sciama, D. W., Waylen, P. C. & Gilman, R. C. (1969), "Generally Covariant Integral Formulation of Einstein's Field Equations", in *Physical Review*, v. 187, p. 1762-1766.
Synge, J. L. (1956), *Relativity: The Special Theory*, Amsterdam, North-Holland.
Tyapkin, A. A. (1972), "Expression of the General Properties of Physical Processes in the Space-Time Metric of the Special Theory of Relativity", in *Soviet Physics Uspekhi*, v. 15, pp. 205-229.
Ushenko, A. P. (1949), "Einstein's Influence on Contemporary Philosophy", in *Albert Einstein: Philosopher-Scientist*, ed. by Schilpp, P. A., The Library of Living Philosophers, La Salle, Evanston & Open Court 1949.
Weiszäcker, C. F. von (1992), *Zeit und Wissen*, München, Hanser Verlag.
Wheeler, J. A. (1964), "Mach's principle as boundary condition for Einstein's equations", in *Gravitation and Relativity*, ed. by Chiu, H. Y. & Hoffman, W. F., Benjamin, New York, pp. 303-349.
Wheeler, J. A. (1988), "Geometrodynamic Steering Principle Reveals the Determiners of Inertia", in *International Journal of Modern Physics* A 3, pp. 2207-2247.
Whitehead, A. N. (1906), *The Axioms of Projective Geometry*, Cambridge, Cambridge University Press.
Whitehead, A. N. (1906), "On Mathematical Concepts of the Material World", in *Philosophical Transactions, Royal Society of London* A, v. 205, pp. 465-525.
Whitehead, A. N. (1907), *The Axioms of Descriptive Geometry*, Cambridge, Cambridge University Press.
Whitehead, A. N. (1915-16), "Space, Time, and Relativity", in *Proceedings of the Aristotelian Society*, v. 16, pp. 104-129.
Whitehead, A. N. (1916), "La Théorie Relationniste de L'Espace", in *Revue de Métaphysique et de Morale*, v. 23, pp. 423-454.
Whitehead, A. N. (1919), *An Enquiry on the Principles of Natural Knowledge*, Cambridge, Cambridge University Press.
Whitehead, A. N. (1919), "Symposium - Time, Space and Material: Are They, and If So in What Sense, the Ultimate Data of Science?", in *Proceedings of Aristotelian Society*, suppl. v. 2, pp. 44-57.

Whitehead, A. N. (1920), *The Concept of Nature*, Cambridge, Cambridge University Press.

Whitehead, A. N. (1920), "Einstein's Theory. An Alternative Suggestion", in *The London Times Educational Supplement*, Feb. 12, p. 83, reprinted in A. N. Whitehead, *Essays in Science and Philosophy*, New York, Philosophical Library 1947.

Whitehead, A. N. (1921-22), "Discussion: The Idealistic Interpretation of Einstein's Theory", in *Proceedings of Aristotelian Society*, v. 22, pp. 130-134.

Whitehead, A. N. (1921-22), "The Philosophical Aspects of the Principle of Relativity", in *Proceedings of Aristotelian Society*, v. 22, pp. 215-223.

Whitehead, A. N. (1922), *The Principle of Relativity with applications to Physical Science*. Cambridge, Cambridge University Press.

Whitehead, A. N. (1929), *Process and Reality. An essay in Cosmology,* New York, The Macmillan Co.

Whitehead, A. N. (1947), *Essays in Science and Philosophy*, New York, Philosophical Library.

Whitehead, A. N. & Russell, B. (1910-1913), *Principia Mathematica*, Cambridge, Cambridge University Press, v. I, 1910, v. II, 1912, v. III, 1913.

Whittaker, E. T. (1953), "The Relativity Theory of Poincaré and Lorentz", in Whittaker, E. T., *History of the Theories of Aether and Electricity. The Modern Theories 1920-1926*, I-II, London, Nelson & Sons, v. 2, pp. 27-77.

Novel “Old Facts”, Old “Novel Facts” and the Periodization as an Epistemological Practice

Flavia Marcacci
University of Urbino
flavia.marcacci@uniurb.it

The logical approach to confirmation renders it so easy that it ceases to be important.

Alan Musgrave (1974: 22)

1. Introduction

The so-called scientific revolution of the seventeenth century involves complex technical elements that require advanced knowledge of both geometry and mathematics and instrumental practice, a deep understanding of historical methods of inquiry, and an appreciation for the philosophical context of the time. Central to such a revolution was the quest for the most accurate astronomical theory among the three fundamental models – that were Ptolemaic, Copernican, and Tychonian ones – which garnered significant attention to many anomalies needed to be solved. On the one hand, astronomical models were developed based on instruments that had many limitations. Only through their gradual improvement could effective solutions be found. Similarly, the evaluation and analysis of data and observations required new mathematical tools. This is why the development of trigonometry, and the introduction of logarithms provided fresh perspectives (Brummelen 2021). On the other hand, evaluating a particular geometric solution often involved a philosophical interpretation, which was either rooted in or challenged ancient concepts, and raised questions about

Flavia Marcacci, “Novel ‘Old Facts’, Old ‘Novel Facts’ and the Periodization as an Epistemological Practice”, in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 197-218

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

ontological references. An additional layer of discussion involved whether ancient natural philosophy could be adapted to new mathematical explanations, or if new mathematical insights would lead to the development of entirely new philosophical categories.

A comprehensive historical investigation is essential to support any interpretation of the developments in modern science. At the same time, the rise of modern science has sparked ongoing, and sometimes conflicting, interpretations of how specific observations, empirical evidence, and theoretical progress led to the eventual dominance of heliocentrism over geocentrism. These debates underscored the need for epistemic clarity and raised broader philosophical issues, such as the demarcation problem and the nature of scientific rationality.

In the 1970s, studies addressing these technical and philosophical dimensions proliferated, conferring the need to clarify the interaction between the history and philosophy of science. The 1970s debuted with the publication of an apical book on the above topic: Alan Musgrave and Imre Lakatos had called eminent scholars to discuss together in 1965 at Cambridge University. The contributions came out in 1970 in the well-known volume *Criticism and the Growth of Knowledge* (Musgrave, Lakatos 1970). Kuhn's proposals and his concept of "normal science" were at the center of the research: all noteworthy speakers – including Karl Popper, Imre Lakatos, Paul Feyerabend – offered direct or indirect criticisms and personal interpretations. Lakatos's ones, proposing a view of how new scientific theories assert themselves over old ones – well known as the *methodology of scientific research programs* – offer here the chance to start our debate.

In the wake of the open discussion, another vital contribution to the relationship between history and the philosophy of science was made just after by Alan Musgrave himself (Musgrave 1974). Another chance was the celebration of the 500th anniversary of Copernicus's birth, inquiring about the Copernican achievement (Westman 1975). Lakatos and Zahar presented a philosophical essay to demonstrate how historical questions deeply resonate within the philosophy of science. Two philosophical issues in science were contextually crucial: the demarcation problem and the existence and nature of universal conditions for a scientific theory.

The first point has significant implications for the history of science, specifically examining the continuous interaction between various areas and forms of knowledge. The second question also impacts the history of science: if such universal conditions exist, then historical forms of science would merely be different expressions of these conditions – a highly controversial hypothesis. Both questions had an echo about the definition of the so-called *novel fact*, conceived as fact able to confirm or disconfirm new scientific theories.

This paper will revisit the key stages of the debate initiated by Lakatos and Elie Zahar, paying particular attention to the responses from historian Neil Thomason. We will discuss how some definitions of novel fact are partial because they do not respect the historical methodology. While this may seem like a dated philosophical discussion, it is still interesting in light of the advancement of historiographical research in recent decades. The problem of the origins of science and its entanglements with the development of a new cosmological view have highlighted how that development was a battle between world systems. But it would be time to understand how there was also a battle between interpretations of that battle.

Far from being able to discuss these competing narratives, this article will emphasize the importance of proper periodizations, considering establishing how good interaction between history and the philosophy of science allows fair philosophical inquiry about science.

2. "Neither Copernicus nor Newton held their own belief", but the revolution took place likewise the same.

Nicolaus Copernicus (1473-1543) inspired generations of historians devoted to studying his astronomy and life. However, a so-called "metacopernicology" has been developing, an investigation of all research ever produced on Copernicus and his writings, viewed through the lens of the history of ideas (Borski-Kokowski 2021). Without definitive sources, multiple interpretations of Copernican proposals continually emerge. It has been said: "we must distinguish between the many unintended consequences his reform turned out to entail and the historical frame in which his efforts make proper sense" (Cohen 2010, p. 106). Interpretations can follow a rigorous historical methodology, where historians clearly distinguish between what the sources reveal and what they infer, ideally acknowledging the historical or theoretical criteria guiding their conclusions. Alternatively, interpretations may be shaped by broader philosophical perspectives, influenced by personal sensitivity or worldview. The first approach generates debates and differing viewpoints, such as the long-standing scholarly focus on Copernicus's Platonic influences[1]. The second approach, however, is tied to specific philosophical theories, where individual cases like Copernicus serve as illustrative examples of broader ideas – such as in the Lakatos-Zahar thesis. As a result, historians may present one version of Copernicus, while philosophers offer another. Ideally, these interpretations would converge into a unified view, but this seems highly unlikely. Finally, the expression "Copernican Revolution" encompasses a vast historical and philosophical metaphor.

Precisely against the ambiguous, or at least uninformative, use of this expression, Lakatos and Zahar denounced that it is not neutral at all because it traced back the astronomical revolution essentially to the publication of *De revolutionibus* (Copernicus 1543). They wrote:

> Let me first define the term 'Copernican Revolution'. Even in the descriptive sense, this term has been ambiguously applied. It is frequently interpreted as the acceptance by the 'general public' of the belief that the Sun, and not the Earth is the center of our planetary system. But neither Copernicus nor Newton held this belief (Lakatos-Zahar 1975: 356).

Lakatos and Zahar wanted to underline how the popular idea of the revolution as a passage from the popular belief in an Earth-centered system to a Sun-centered one falls outside of the history of science in a strict sense. In their opinion, the Copernican revolution must be circumscribed exactly to this statement: "the hypothesis that it is the Earth that is moving around the Sun rather than vice versa, or, more precisely, that the fixed frame of

---

[1] The idea that Platonism and Neoplatonism influenced Copernicus' astronomy dates back to scholars such as Burtt 1959, Koyré 2009, and Kuhn 2003, to name just a few. In engaging with this notion, Rosen (1983) questioned whether such attributions were accurate, noting they were based on only a few references in Copernicus' work. Interestingly, Kepler traced a line of intellectual descent from Plato to Copernicus (Eastwood 1982). While this suggestion remains largely speculative, it has found some support in broader historical analyses that link Platonic thought to Ptolemaic astronomy (Gingerich 2002). Copernicus' elimination of the equant – often viewed as anti-Ptolemaic and pro-Platonic – and his emphasis on harmony and the regularity of the cosmos, seem to give his astronomical program a certain philosophical coherence (De Pace 2009; Vesel 2014). However, implicit references to Platonic thought should be carefully distinguished from practical astronomical reasoning (Hatfield 1990). The former may be hypothesized; the latter must be assessed on firmer grounds. What we can say with confidence is that *De revolutionibus orbium coelestium* was primarily intended to communicate with fellow astronomers, not philosophers, even though it undoubtedly reflects some philosophical underpinnings.

reference for planetary motion is the fixed stars and not the Earth" (Lakatos-Zahar 1975: 356-7).

Philosophers had read nonetheless the origin and development of the Copernican revolution provided in some criterion of scientific rationality and had proposed several key interpretations, all of which rejected by Lakatos and Zahar. What follows is a focused summary of their synthesis of these approaches.

1. Assumption of a demarcation criterion for scientific rationality.
    a. Empiricist accounts for the Copernican revolution. The core of this proposal is that the superiority of Copernican hypothesis is due to its straightforward empirical base. In this group, one counts:
        i. *strict inductivists*: Copernicus deduced heliocentrism from the facts. Robert Bellarmine and Karl Popper later criticized this approach, but even now, some believe it is plausible. Above all, geocentrism was in accordance with the facts.
        ii. *probabilistic inductivists*: the best scientific theory had a major probability to deduce laws from the facts, in the manner of a bayesian interpretation of the Copernican revolution. Nevertheless, nobody succeeded in confirming the supposed superiority.
        iii. Falsificationists organized in two groups following two kinds of falsificationism: (1) Based on the idea that every Ptolemaic effort is an *ad hoc* accommodation of the theory. However, Owen Gingerich has demonstrated that the Alphonsyne tables were founded on the *single-epicycle* technique and not *epicycles on epicycles*; therefore, one must clarify what really was an *ad hoc* accommodation. (2) Based on the idea that both Ptolemaic theory and Copernican theory were *refutable* for a long time until a *crucial experiment* denied the first and held the second. However, in such a case, the crucial experiment would have occurred outright in 1838, when Bessell discovered the parallax effect, and not in 1723 with the discovery of the aberration of light deprived by a thorough explanation.

   b. "Semplicist" – so called by Lakatos and Zahar those conventionalists who adopt some criterion to choose one theory among others and to avoid relativism[2] – accounts for the Copernican revolution. The empirical ground is often insufficient to justify the affirmation of a scientific theory. Conversely, conventionalist approaches deem each scientific theory adaptable to every context because it is fundamentally built on a series of conventions. A strict conventionalist approach leads to Relativism which makes Ptolemaic and Copernican theories substantially equivalent. A simplicist reaction wants to avoid any form of Relativism, and the choice of a theory is based on other criteria – for instance, a more coherent or more straightforward shape (*superempirical virtue*). Accordingly, Copernicus was the first to claim his theory was simpler than Ptolemy's one. However, some scholars complained about some complications in the Copernicus program, making that theory difficult and not beautiful.

2. Absence of a demarcation criterion for scientific rationality.
   a. Elitism. Only *case law* exists, not *statute law* (as according to Polanyi). Heliocentrism prevailed as an inarticulable *fingerspitzengefuhl* owned by the elite, who chose her favorite theory. However, if so, why did so few astronomers follow Copernicus before Kepler and Galileo, whereas more astronomers did it later?
   b. Relativism. Only the best propoganda wins. And also, Copernicanism became metaphysics (as according to Feyerabend).

Opinion 2.b is the strongest and most challenging to deal with. As a further interpretation, Lakatos proposed the *methodology of scientific research programs* as a new demarcation theory. It is made of a hard core, heuristic, and protective belt. Each theory can be theoretically (if each modification leads to new unexpected predictions) or empirically (in the case of corroboration of some novel predictions) progressive. Ad hoc manoeuvres degenerate the program. But no research program solves all its anomalies. Instead, it lives with them. In any case, it contains a heuristic advance. A

[2] Lakatos-Zahar 1976, p. 361: "I use this rather ugly term for methodologies according to which one cannot decide between theories on empirical ground: a theory is better than another if it is simpler, more "coherent," more "economical" than its rival".

program is superseded when the new program predicts everything the old one predicted plus more. Lakatos discusses the problem of empirical equivalence: two equivalent theories can be evaluated not only in terms of their falsifiability and evidence, but also in relation to their heuristic capacity and as a function of temporal factors. According to Lakatos, Copernicus' program was *theoretically* progressive and capable of anticipating facts, so called *novel facts*.

> I originally defined a prediction as "novel", "stunning", or "dramatic" if it was inconsistent with previous expectations, unchallenged background knowledge and, in particular, if the predicted fact was forbidden by the rival program (Lakatos-Zahar 1976: 375).

In this case, the observation of Venus' phases was like a novel fact but happened only in 1610 by Galileo so that until that time Copernican program was not progressive empirically and astronomers had no rational criterion to choose it.

Zahar modified the methodology of scientific research programs by reinforcing the criterion of evidence and took the case of Mercury's anomalous perihelion. The problem of incongruency in Mercury's orbit has been well-known for almost a hundred years to the point that no new observations were needed: it needed new explanations. It was Einstein to furnish it by means of the General Relativity Theory (Zahar 1973.2). The debate shifted on the definition of what is new and what is a *novelty*.

> My re-definition of novelty amounts to the claim that in order to assess the relation between theories and empirical data within a research-programme, one has to take into account the way in which a theory is built and the problems it was designed to solve (Zahar 1973.1, p. 103).

According to Zahar, a novel fact must not play any role in the construction of a theory. Indeed, a novel fact à la Zahar supports a new theory when it is already known and makes the fact explainable because of new theoretical assumptions. Copernicus assumed as fundamental the different elongation of superior and inferior planets (Venus and Mercury on one hand, Mars, Jupiter, and Saturn on the other hand). From this fact, he *explained* – and then *predicted* – other known facts with an easier explanation (see Table 1, column 1): stations and retrogressions of planetary motion, different periods of superior planets (=Mars, Jupiter, Saturn) and inferior ones (= Mercury, Venus), explanations of many discrepancies from the various distance from Sun, and many others. Therefore, the Copernican proposal was progressive.

In order to reinforce this interpretation, Lakatos and Zahar set up an *historical thought-experiment*. Going back to the 1520 or earlier, certain retrogrades and stations of the planets and the Sun around the Earth had never been experimentally observed by two astronomers for cause as fog or cloudy skies. Until future observations, they could have believed in two different systems, for instance, based on circular planetary motions and Sun-worship for a Copernican and on circular planetary motion justified by deferent end epicycle with the Earth at rest for a Geocentrist. They easily could have coexisted and observed the same facts. Thus, the superiority of Copernicus's system would prove itself later, at the time of new observations, even if it must have been in nuce even earlier.

The historian Thomason attacked this historical mental experiment and judged it as a case of *fictional history*. However, real history was not the fictional one proposed by Lakatos and Zahar. Thus, Thomason corrected many vagueness and historical uncertainties in Zahar's proposal and demonstrated that the Zahar criterion for novel facts determined too many fictional facts which distort the real story: "if a research programme is assessed on the basis of fictional history, it will appear to have many more 'novel facts' than its real history entitles it to" (Thomason 1992: 191). If Zahar's criteria had been effective, historical records would have been respected.

Thomason's indictment goes in two main directions.

1. Methodological. What sense does fictional history have? Often the real history contains so many variables that it could also include fictional history, but it could have even more. The relation between real and fictional is always a risk to judge the real history.
2. Content-related. Even admitting fictional history and assuming something as a modified Zahar's criterion, the examples in support of the Lakatos and Zahar' thesis do not hold. According to Thomason, Zahar's definition of novel fact reduces the historical inquiry to the following: "one must answer a historical question: whether that fact played a role in the scientist's construction of the theory or the problem it was designed to solve?" (Thomason 1992: 163). If the fact did not play a role in the construction of the theory, then it is a novel fact when it receives a new explanation.

Thomason focuses on some elements, especially on the existence of many geocentric and heliocentric proto-theories. Thus, not only Ptolemaic, Copernican and Tychonian models were available for the ancient astronomers, but also other world-systems and theories, which gathered and sometimes mixed-up elements one from the others. Some alternatives to methods and data from Ptolemy or Copernicus were, for instance, from Apollonius and Aristarchus. The fictional history could consider more options, like those from pseudo-Eudoxian system. Thomason follows with a painstaking analysis to distinguish true history and fictional history and evaluate Zahar's criterion. The result is discouraging: true history counts very few cases of novel facts, against fictional history, which accumulates a large amount of them (Table 1).

	Fictional History				Real History				
	proto-Eudox.	proto-Ptol.	proto-Coper.	proto-Tycho	Eudox.	Apoll./Ptol.	Aristarch.	Coper.	Tycho
(1) Retrogradations	N	Y	Y	Y	N	N	N	N	N
(2) Brightness	N	Y	Y	Y	N	N	N	N	N
(3) Resolved Variations	N	N	N	N	N	N	N	N	N
(4) Inconstant Periods	N	Y	Y	Y	N	N	N	N	N
(5) Solar Component	N	Y	Y	Y	N	N	?	N	N
(6) *Superior Retrogrations*	N	Y	Y	Y	N	N	?	N	N
(7) Inferior Retrogrations	N	Y	Y	Y	N	N	?	N	N
(8) Period v. Distance	N	N	N	N	N	N	N	N	N
(9) Determining Distance	N	N	N	N	N	N	N	N	N
(10) Retrograde Arc Length	N	N	Y	Y	N	N	Y	Y	N
(11) Bright Retrograde	N	Y	Y	Y	N	N	?	N	N
(12) Moon's Orbit	Y	N	N	N	N	N	N	N	N
(13) Stars Don't Brighten	Y	Y	N	Y	?	N	N	N	N
(14) No Stellar Parallax	Y	Y	N	Y	N	N	N	N	N
(15) South Celestial Pole	Y	Y	Y	Y	Y	N	N	N	N
Number of Novel Facts 'Y'	4	10	19	11	1	0	1	1	0
Number of Possibly Novel Facts '?'	0	0	0	0	1	0	4	0	0

Table 1. Combinations between facts and hypotheses in real and fictional history according to Thomason. The table indicates if each fact is explained or not in the various theories. In the bottom rows, the sum of new facts for each theory is indicated.

3. Searching for a "logico-historical approach to confirmation": novel facts and time-ordered facts.

Musgrave discussed the problem of the relation between history and philosophy in science focusing well how "explaining known facts is one thing, predicting new facts another" (Musgrave 1974: 2). Consequently, if a predicted fact happened, it becomes a confirmation or not of a theory. Musgrave defines the *purely logical (or logical, for short) approach to confirmation*: a fact is considered as a *piece of evidence e* in favor of a hypothesis *h* which explains or predicts it, and one must consider the logical relation between them. "It is quite irrelevant whether *e* was known first and *h* proposed to explain it, or whether *e* resulted from testing predictions drawn from h" (Musgrave 1974: 2). The bond between evidence and conjecture is so fundamental that it "cannot depend upon whether the evidence came to be known before the theory was proposed or afterward" (Musgrave 1974: 3). Musgrave stresses that the logical approach to confirmation has produced some paradox and complication, above for having omitted that value – at least, some form of respect – of intuition linking that evidence and the hypotheses. For that, Musgrave admits the necessity of some historical ingredient in confirmation and supposes something as a *logico-historical approach to confirmation*. Musgrave continues by exploring the background knowledge of any theory and its heuristic value.

> All variants of the historical approach will make the confirmation of a scientific theory somehow depend upon the historical setting in which that theory was proposed. Of course, once the actual content of background knowledge has been ascertained by historical investigation, the analysis of confirmation proceeds logically. But we investigate the logical relations between three things (theory, evidence, and background knowledge) and not two as in the purely logical approach (Musgrave 1974, p. 7).

Since we have a "strictly temporal view of background knowledge", then all facts known before the proposal of a hypothesis are not valid to confirm or lessen the hypothesis itself. Moreover, the concept of background knowledge remains foggy and susceptible to subjectivistic and relativistic interpretations that are unlikely to have logical value. Musgrave claims that a strictly temporal background knowledge helps to distinguish hypotheses based or already known facts from <u>those considerable as predictive of novel facts</u>.

After a long discussion, Musgrave's purpose is clear: to make "historical approach to confirmation a little more palatable" (Musgrave 1974: 21).

According to Worrall (2006: 31):

> Although Alan's paper was published in 1974, the problem it faces has not been given a satisfactory resolution – at least not one that has met widespread acceptance. It remains very much a live issue within current philosophy of science.

The debate took the wide path, especially in John Worrall's work but also in Zahar's one, toward the question of the type of scientific realism and empirical evidence for grasping the need for the trade-off between explanatory power and descriptive accuracy. Worrall particularly proposed structural realism for putting together the best of antirealism criticisms and realism accounts (Worrall 1989). Nevertheless, this approach stands on a pure logical level and – at least as I understand it – it does not make any "historical approach to confirmation a little more palatable".

In other words, the debate about theoretical and empirical contents in science was conducted as if such distinction was without ambiguity and, above all, without considering historical analyses, which often call into doubt exactly such distinction. Besides, historical methods can improve logical inspection, a fact that is ignored. On the contrary, historical considerations to clarify how empirical data and concepts or ideas were synthesized, as well as recognize how some historical tools can support logic, are fundamental to making philosophical recommendations more suitable and valid in the light of actual history.

In general, all scholars agree that any assessment of novel facts should be returned to the original context. As Lakatos and Zahar with the above-quoted definitions, each scholar connects the novelty with the possibility of either explaining or previewing the fact itself. Mainly, Lakatos has spoken about "previous expectations", "unchallenged background knowledge", and fact which "was forbidden by the rival program". So Lakatos acquires a way to judge theories in history and states that (a research programme is) theoretically progressive if each modification leads to new unexpected predictions and it is empirically progressive if at least some of these novel predictions are corroborated (Lakatos-Zahar 1976: 369).

Zahar tried to introduce some historical dimension when conducting a novel fact to its context, too. He affirms:

> A fact will be considered novel with respect to a given hypothesis if it did not belong to the problem-situation which governed the construction of the hypothesis (Zahar 1973.1: 103).

For that any fact must be collocated in its original context. However, he finishes distorting the real history, as Thomason denounces. The historian rejects Zahar's criterion because it can be applied only to fictional history. Indeed, Thomason argues that fictional histories contain a much higher number of novel facts than real historical accounts. He points out that while Lakatos and Zahar regarded Copernicus as superior to Ptolemy and Brahe, only a small group of astronomers accepted Copernicus' system during the period between *De revolutionibus* and Galileo's observations. Consequently, it seems like if most astronomers should have been irrational.

He also criticizes Zahar's criterion, because no criterion is helpful when historical data are insufficient to reconstruct events accurately. Moreover, Thomason highlights that an old fact can serve both as an inspiration and support for a scientific theory, in a way which is incompatible with Zahar's concept of novel facts. The method used to construct a theory doesn't undermine the value of the theory itself: for instance, Tycho was guided by some Copernican insights, as he was also reconciled by a geo-heliocentric idea of the world when he realized that comets debunked solid spheres and in his model of the world could intertwine the orbits of the Sun and Mars (Thomason 1992: 179-180).

> [I] assume simply that Brahe developed an original research programme and that it was guided in part by the standard Ptolemaic and Copernican ways of predicting planetary orbits, by these facts which were seen as relevant to determining the nature of the cosmos, and by the explanatory power of Copernican theory (Thomason 1992: 181).

Thomason concludes that, according to Zahar's criterion, much of science during history would appear irrational, and the specific details that would qualify certain facts as novel to Zahar are irrelevant to evaluating the overall quality of a theory.

Musgrave also traced back science to its historical contextualization but followed the road of the difference between something testable and something already known but not testable.

> …a theory is independently testable only if it predicts a *novel fact*, a fact not 'known to science' when the theory was proposed (Musgrave 1974: 15-6).

During these decades and more recently, John Worral (Worrall 2006, 2008, 2014) has been working on this issue and affirmed:

> The issue of *prediction vs. accommodation* is a long-running one that continues to be hotly debated. There seem, however, to be two obvious problems with the suggestion that predictions carry more supportive weight than explanations of (otherwise equivalent but) already established facts. The first is that while the suggestion yields the intuitively correct judgments in some cases, *it does not do so in all*. The facts about the precession of Mercury's perihelion were, for example, well known before the general theory of relativity was articulated, and yet all serious commentators regard that theory's explanation of Mercury's orbit as constituting important empirical support for it – at least as strong support as it received from the prediction of any temporally novel fact. The *second problem* is more general: the suggestion seems to stand without any epistemic justification –

> why on earth should the *time-order* of theory and evidence have any epistemological import? (Worrall 2008: 285, *italic* does not reproduce the original).

According to Worrall, the first problem notices how the facts that make a theory grow are sometimes not new. But above all the second point seems address to history a special role, insofar as history is the analysis and reflection on what I would call *time-ordered facts*. For this reason, Zahar's criterion furnishes a good indication but is insufficient. If it is right to argue that new facts should be understood within their context and the background knowledge that makes them "novel", the role of temporal order cannot be overlooked. The time-line useful for the framework inside which accepting or rejecting a hypothesis must be pointed out, discussed, even refuted but only after an evaluation. Musgrave and maybe Lakatos before tried to introduce a similar idea when they aim to stress how timeline is crucial to understand a fact as novel or not (Musgrave 1974). Worral has gone recently back on this question, particularly debating how time-order counts in a scientific evaluation of a theory (Worrall 2014). At the center of the discussion, again, remains the question of how a piece of evidence, known or not before or after the confirmation of a theory, is involved in it. The time-order influences or not the epistemic evaluation of the theory, depending on the opinions of the scholars. However, the judgment about time-order, a historian's typical activity, does not interest philosophers. Or, at least, it is totally evaded by the discussion. It could be interesting to understand how novel facts interact or may be considered in the light of the time-ordered facts. However, time-ordered facts result from historical activities and judgments, precisely called "periodizing". This is one of the most challenging operations on the historian's table. Something practical or methodological may have exciting implications in developing philosophical issues.

4. *Novel old facts* and *old novel facts*: what was supposed to happen, it happened.

The above discussion did not bring convergent results (Nugayev 2013). Thomason outlines his counterexamples to Lakatos-Zahar in detail. However, how and if history can help philosophers better focus on their issues remains outstanding. Besides, Thomason forgoes establishing at least a more general criterion of comparison (if not also a criterion of choice) between *old* and *new* evidence for a scientific theory.

Two aspects of the discussion remain unsolved up to this point.

1. First, how mixing up what is old and what is new in a scientific theory. Is time-ordered reconstruction enough to understand it?
2. Second, if something in the historian's (not fictional) work is valuable to elaborate epistemic and logical solutions for understanding how science actually works. What historiographical tool best defines the logical and epistemic problems posed by both philosophers and historians?

In order to find answers to both questions, I will try to show how a good periodization can safeguard the need to establish criteria of rationality at least in the case-study considered in this discussion, that is the Copernican revolution.

The long period 1543, corresponding to the publication of Copernicus's *De revolutionibus*, to 1687, corresponding to the publication of Newton's *Philosophiae Naturalis Principia Mathematica* on universal gravitation, is generally considered the historical parenthesis for the development of the new science. About the beginning, as we said, Lakatos and Zahar polemized and alerted about the risk to forget how long the spreading of Copernicanism (Westman 2020) was. Their complaint stood at the beginning of decades of discussion. Indeed, the historiographical concept of the "scientific revolution", rooted in the concept of Copernican revolution, has recently been problematized (Nickles 2009; Schlagel 2015; Wray 2024; Omodeo 2020). Many works multiplied historiographical perspectives (Cohen 2011; Daston 2017; Henri 1997; Renn 2020), discussing technical aspects in the transformation of mathematical and instrumental astronomy (Linton 2004), distinguishing astronomy and mechanics from natural science, medicine or biology for which the developments followed other paths (Kelly 2010, Clericuzio 2022). The theme is so extensive that it is impossible resuming it in a few lines, so much to the point that our inquiry about what were the novel facts in the scientific revolutions seem just a lucubration. However, our inquiry investigates the possibility of creating room for interaction between history and philosophy and for understanding science as a historical enterprise constituted by rational criteria that succeeded along the ages. To give order and historical coherence to such criteria, it is fundamental that a philosopher situate facts and their interpretation in the correct backdrop.

Going back to the Copernican Revolution, one should also highlight risks about the above ending limit. Newton furnished the most advanced reformed Copernican system, but so many elements changed that it is questionable if scientific rationality emerged for a few causes (Buchwald and Feingold 2012). Lakatos and Zahar problem was why Copernican model superseded

Ptolemaic one. In such a problem, it is mandatory considers that the long period Copernicus-Newton counts at least two sub-periods:

1. 1543 to 1609, from the publication of *De revolutionibus* to Galileo's telescopic observations of many celestial novelties.

2. 1609 to 1687 when the telescopic discovery by Galileo, above all Venus's phases in 1610, definitively defeated the Ptolemaic system, Kepler defined the planetary laws, and Newton founded them on the gravitation law.

This division facilitates to differentiate the problems of astronomy. Indeed, some problems are present in both phases: not only the possibility of terrestrial motions but also, for instance, the determination of the solar and the lunar motion was studied continuously, as well as the precession of the equinoxes or the evaluation of the best measurement unit. However, circumstances provoked astronomers' agenda in such a way that every problem was oriented to different ends. The first stage offered the scenario for two main competitors, Ptolemy and Copernicus, flanked by Tycho around upon the 80s of the sixteenth century (Brahe 1588, 1610). Tychonian system admitted the revolutions of the Sun around the Earth and all the other planets around the Sun. The technical context was that of naked-eye astronomy. Tycho perfected the astronomical tables and improved the observations, but he did not adhere to Copernicanism for lack of some explanations. His intermediate system successfully determined the motions of the planets, particularly the inner ones. Unlike Copernicus, who proposed excessively large distances for Saturn and the fixed stars, along with exaggerated stellar diameters, he avoided such overestimation. Additionally, his system didn't require proofs of Earth's motions, whereas the Copernican model heavily relied on them but lacked such evidence (Small 1804). In this first period, astronomy maintained its astrometric vocation, setting some epistemological issues very close to ancient astronomy, for instance, the separation between celestial and terrestrial physics.

In the second stage, the struggle became between heliocentrism and geo-heliocentrism, between a Copernican system modified by Keplerian elliptical orbits against Tychonian and Tycho-derived world-systems. In such a contrast, searching for observations and pieces of evidence, as well as elaborating new physical principles and deductive demonstrations became the most important objective. For example, the question of the composition of the heavens was a central issue in the early stages. The nature of comets, particularly whether their paths lay above or below the lunar sphere, was a key astronomical topic, raising astrological questions and attracting prominent astronomers. Tycho Brahe gained fame by proving the fluidity of the heavens. However, in later periods, this topic became less crucial, though debates like that between Orazio Grassi and Galileo (Galileo 1623) still influenced scientific methodology, for instance highlighting how to interpret data, the role of doubt in inquiry, and the unfair weight of authority in the argumentation (Chappell 2024). With the telescope's use, focus shifted to understanding optical laws, essential for determining the magnitude, not just the proportionality, between the sizes of planetary orbits taking advantage of the Keplerian laws.

As astronomical problems and practices evolved, so did the related epistemic questions. Here are a few examples. The ancient role of astronomy in "saving the phenomena" increasingly revealed its ambiguities. In the first phase, Copernicus' system introduced a new geometric model that challenged the Aristotelian view of nature. Aligning demonstrations with data became a pressing issue: astronomy continued to save the phenomena through hypotheses, but these hypotheses were increasingly grounded in the *reality* of the heavens, not merely in their *appearances*. Consequently, clarifying the epistemic role of astronomical knowledge became essential. Positional astronomy could justify different models, but it needed to be paired with a physics capable of providing a foundation for mathematics. In the second phase of the debate, the telescope revealed new aspects of the sky, leading to similar consequences but introducing a new problem: what is the relationship between what we perceive and the objects that the philosopher of nature studies? (Camerota, Giudice 2023).

Thomason, Lakatos, Zahar and other scholars always mentioned the importance of the context but underestimated how it affects the centrality of one issue or another. This cannot fail to have consequences also for a novel fact. Take back on the Venus' phases case-study, quoted in the debate. Such a fact was Zahar-novel because the possibility of Venus's phases was known and was attended to evaluate if the planet would rotate around the Earth or the Sun. Only a complete cycle of phases – and not a partial one – is compatible with a Sun-centered system, and for that it was crucial its observation (Thomason 1994: 327; Palmieri 2001: 114-116). Nevertheless, it was not a novel fact *a là* Zahar because it was only an empirical hypothesis to be observed: it needed new instruments to be observed and verified, and not new categories to be explained and understood. New categories for understanding this phenomenon would be developed in earnest after his observation.

Thus, we could speak about *novel "old facts"* and *old "novel facts"*. In the former case, *novel "old facts"*, a fact is old because already verified and studied, and it is new because the categories for its understanding must change. For instance, retrograde motion for a planet received a totally new explanation by mean of Copernican system, even if it was well-known. In the latter case, *old "novel facts"*, a fact can be expected from long past time but not yet observed as long as it happens to observe it. For old new facts, what was supposed to happen, it happened. For example, in the Ptolemaic system, Venus' phases were expected to appear either sometimes below or above the Sun, without a full phase followed by a crescent. In contrast, the Copernican system predicted a complete cycle of phases – similar to the Moon, though not identical (Palmieri 2001). When Galileo observed a fully illuminated Venus followed by a crescent phase, he became increasingly convinced of the Copernican system. This was possible because of Galileo's telescope. However, the hypothesis of Venus' motion around the Sun was ancient, and recently it was present also in Tycho. Venus' phases discriminated between the Ptolemaic and Copernican hypotheses, rejecting the first and confirming the second (Gingerich 2011). Yet, this *old novel fact* could not determine alone if Copernicus or Tycho was right. Besides, no single fact could have provoked the abandonment of the geocentric theory. The battle involved many more elements, from outside and inside the theories, from mathematical and instrumental traditions, from astrometric and physics. Its upper temporal boundary coincided with Newton's theory of gravitation, which terminated the dispute and established the winner. Unfortunately, the final discrimination needed other empirical evidence, which had been coming for a long time. Indeed, scientific instruments had to be improved and theories to become more refined: so, stellar parallax and the Earth's rotation – facts valuable in demonstrating respectively the revolution and the rotation of the Earth – were validated during the centuries XVII-XVIII. To be further precise, besides, observing the phases of Venus also raised questions about the nature of the planet and how it interacts with light (Thomason 1994). A last remarkable feature of the winning heliocentric theory is that it was no longer the original Copernican version, but a new version made possible by new conceptual and mathematical tools.

On the cover of these considerations, as Thomason (1992) argues, it is fair to avoid fictional history in the context of philosophical debate, but also more in general, if the end of an account is understanding what really happened. At the same time, historical developments of science support rightly philosophical argumentations and they must do it. A first and necessary compromise is the introduction of good periodization, referring any problems to its exact context. Indeed, if scientific problems changed alongside its history, then the logical evaluation of such contents must weigh the change. Old new facts are bearer of novelty because they need new technology and a more complex theoretical framework to be confirmed and understood.

5. Conclusion: Can History of Science say something epistemic to Philosophy of Science?

The long debate over novel facts has often overlooked historical accuracy. While references to Tycho are common, they frequently lack substantial value. In recent decades new historical analyses have added many elements, demonstrating how a philosophical discussion without robust historical references risks superficial evaluations or even errors. This paper offers a reassessment grounded in a methodological principle that all participants in the debate implicitly supported but never fully addressed: the importance of careful historical periodization. While historians often invoke periodization to critique philosophical accounts, it has rarely been employed as a constructive epistemological tool – particularly in clarifying the concept of the "novel fact." By distinguishing between pre- and post-Galilean contexts, this paper argues that historical framing allows us to identify two distinct categories: *novel "old facts"*– known phenomena explained with new interpretations – and *old "novel facts"* – hypothetical phenomena later confirmed through new technologies. The phases of Venus as an *old "novel fact"* serve as a compelling case study to illustrate how this nuanced approach deepens our understanding of scientific discovery.

Recognizing the overgeneralizations and pluralism in defining a novel fact raises questions about whether historiographic practice can genuinely support philosophical inquiry, as it often complicates matters. Nevertheless, historiography provides valuable insights in several ways:

1. *Distinguishing novel facts from temporally ordered facts*. Understanding novelty depends on various factors, including the original theoretical context and the meaningfulness of certain assumptions. Thus, defining a novel fact requires not only logical confirmation but also *historical confirmation*.
2. *Avoiding fictionalized histories*. It is crucial to circumscribe and minimize the risks of fictionalized history. While hypothesizing about facts can be useful, exaggerating their significance in philosophical

and historical discussions is misguided. The history of science must rely on reliable sources and data to prevent misinterpretations and oversimplifications of the origins and development of scientific theories.

3. *Adjusting epistemic evaluations to the historical context*. The questions raised by Lakatos and Zahar have been pivotal for generations of philosophers. However, considering these questions in the context before and after the telescope alters the definition of novel facts, allowing for a more nuanced understanding of the many facets of complex astronomical problems. Novelty can arise for various reasons, each as important as analytical definitions.

In light of this, the question becomes very deep: what is scientific rationality if theories evolve over time? Throughout history, science has continuously sought theories that adequately describe natural phenomena. Yet history itself is marked by constant change (Marcacci 2023), revealing that science has exhibited multiple forms of rationality. On one hand, the history of science can provide valuable epistemic insights into the philosophy of science, highlighting its historical nature. On the other hand, historical analysis risks getting lost in detail unless it embraces theoretical perspectives inspired by the philosophy of science. For now, we can accept that scientific rationality exists not *despite* its history, but *because* of it. The “logico-historical approach to confirmation” desired by Musgrave should move from this path.

Acknowledgements

I wish to express sincere gratitude to this volume’s editors Claudio Ternullo and Matteo Antonelli for their patient and careful editorial oversight and to Vincenzo Fano for his thoughtful recommendations. I am also indebted to the anonymous reviewers for their insightful critiques, which contributed substantially to improving the original manuscript. Any remaining inaccuracies or mistakes are entirely my responsibility.

References

Borski, G. & Kokowski, M. (2021), “Copernicus, his Latin style and comments to Commentariolus”, in *Studia Historiae Scientiarum* 20 (September), 339-438.

Brahe, T. (1588), *De mundi aetherei recentioribus phaenomenis*, Uraniburgi.

Brahe, T. (1610), *Astronomiae instauratae progymnasmata*, Francoforte.

Brummelen, G. Van (2021), *The Doctrine of Triangles A History of Modern Trigonometry*, Princeton, NJ, and Oxford, Princeton University Press.

Buchwald, J. & Feingold, M. (2012), *Newton and the Origin of Civilization*, Princeton (NJ), Princeton University Press.

Burtt E. A. (1959), *The Metaphysical Foundations of Modern Physical Science*, London-Bradford, Lund Humpries (first edition 1924, second edition revised 1932).

Camerota, M. & Giudice, F. (2023), "Introduzione", in G. Galilei, *Il Saggiatore*, Milano, Hoepli.

Chappell, E. (2024), "A world of words: Rereading Galileo's grand book of philosophy from Il Saggiatore". Galilæana XXI, 1 (2024): 53-80

Clericuzio, A. (2022), *Uomo e natura. Scienza, tecnica e società dall'antichità all'età moderna*, Roma, Carocci.

Cohen, H. F. (2010). *How Modern Science Came into the World: Four Civilizations, One 17th-Century Breakthrough*. Amsterdam University Press.

Cohen, H. F. (2011), *The scientific revolution: a historiographical inquiry*, Chicago, The University of Chicago Press.

Copernicus, N. (1543), *De Revolutionibus orbium cœlestium libri VI*, apud Jo. Petreium, Norimbergae.

Daston, L. (2017), "The history of science and the history of knowledge", in *Know: A journal on the formation of knowledge*, 1/1, pp. 131–154.

De Pace, A. (2009), *Niccolò Copernico e la fondazione del cosmo eliocentrico. Con testo, traduzione e commentario del Libro I de Le rivoluzioni celesti,* Milano, Bruno Mondadori.

Eastwood, B. S. (1982). Kepler as Historian of Science: Precursors of Copernican Heliocentrism according to "De Revolutionibus," I, 10. *Proceedings of the American Philosophical Society*, *126*(5), 367–394. http://www.jstor.org/stable/986213

Gingerich, O. (2002), "The Trouble with Ptolemy." *Isis*, vol. 93, no. 1, 2002, pp. 70–74. *JSTOR*, https://doi.org/10.1086/343243.

Gingerich, O. (2011), "Galileo, the telescope, and the modern astronomy", in *Proceedings of the American Philosophical Society*, 155.2 (June), pp. 134-41.

Hatfield, G. (1990), "Metaphysics and the new science", in David C. Lindberg & Robert S. Westman (eds.), *Reappraisals of the Scientific Revolution*, Cambridge, Cambridge University Press. pp. 93–166.

Henry, J. (1997), *The Scientific Revolution and the Origins of Modern Science*, London, Red Globe Press.

Henry, J. (2010), *The scientific revolution and the origins of modern science* (3. ed., reprint). Palgrave Macmillan.

Koyré, A. (2009), *The Astronomical Revolution: Copernicus, Kepler, Borelli*, Routledge, London, 2009 (ebook, first published 1961).

Kuhn, T. (2003), *The Copernican revolution: planetary astronomy in the development of western thought*, Harvard University Press, Cambridge, MA, 2003 (first published 1957).

Lakatos, I. & Zahar, E. G. (1975), "Why Did Copernicus's Program Supersede Ptolemy's?", in R. Westman (ed.), *The Copernican Achievement*, Los Angeles, University of California Press, pp. 354–83.

Lakatos, I., & Musgrave, A. (1970), *Criticism and the Growth of Knowledge, Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*, Cambridge, Cambridge University Press.

Linton, C. M. (2004), *From Eudoxus to Einstein A History of Mathematical Astronomy*, Cambridge, Cambridge University Press.

Marcacci, F. (2023), "Scientific Change, Realism, and History of Science", in A. Angelucci, V. Fano, P. Graziani, G. Galli, & G. Tarozzi (eds.), *Scientific Change, Realism, and History of Science. Festschrift for Mario Alai*, Milano, Franco Angeli, pp. 61-71.

Musgrave, A. (1974), "Logical versus Historical Theories of Confirmation", in *The British Journal for the Philosophy of Science*, 25.1 (Mar.), pp. 1-23.

Newton, I. (1687), *Philosophiae Naturalis Principia Mathematica*, London, Societas Regiae ac typis Josephi.

Nickles, T. (2024), "Scientific Revolutions", in *The Stanford Encyclopedia of Philosophy* (Spring Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/spr2024/entries/scientific-revolutions/

Nugayev, Rinat M. (2013), "The Ptolemy-Copernicus transition: Intertheoretic Context", in Almagest 4(1), pp. 97-117.

Omodeo, P D. (2020), "Ideologies of the scientific revolution: the rise and fall of a historiographical concept", in Jalobeanu, D., and Wolfe, C.T. (eds), *Encyclopedia of Early Modern Philosophy and the Sciences*, Cham, Springer.

Palmieri, P. (2001), "Galileo and the Discovery of the Phases of Venus", *Journal for the History of Astronomy*, 32(2), pp. 109-29. https://doi.org/10.1177/002182860103200202

Renn, J. (2020), The Evolution of Knowledge: Rethinking Science for the Anthropocene, Princeton, Princeton University Press.

Rosen, E. "Was Copernicus a Neoplatonist?" *Journal of the History of Ideas*, vol. 44, no. 4, 1983, pp. 667–69. *JSTOR*, https://doi.org/10.2307/2709222. Accessed 15 Apr. 2025.

Schlagel, R. H. (2015). *Three scientific revolutions: how they transformed our conceptions of reality*. Humanity Books.

Small, R. (1804), *An Account of the Astronomical Discoveries of Kepler*, London, Mawman.

Thomason, N. (1992), "Could Lakatos, even with Zahar's Criterion for novel fact, evaluate the Copernican Research Programme?", in *British Journal for the Philosophy of Science*, 161-200.

Thomason, N. (1994), "Sherlock Holmes, Galileo, and the Missing History of Science", in *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Volume One: Contributed Papers, pp. 323-333.

Vesel, M. (2014). *Copernicus: Platonist Astronomer-Philosopher: Cosmic Order, the Movement of the Earth, and the Scientific Revolution,* Frankfurt am Main: Peter Lang.

Westman, R. (ed.) (1975) *The Copernican Achievement*. Berkeley-Los Angeles-London: University of California Press.

Westman, R. S. (2020), *The Copernican Question. Prognostication, Skepticism, and Celestial Order*, University of California Press (1st ed. 2011).

Worrall, J. (2006), "Theory-Confirmation and History", in C. Cheyne & J. Worrall (eds.), *Rationality and Reality: Conversations with Alan Musgrave,* 31-61.

Worrall, J. (2008), "Theory-Change in Science", in S. Psillos and M. Curd (eds.), *The Routledge Companion to Philosophy of Science*, London-New York, Routledge, pp. 281-91.

Worrall, J. (2014), *Prediction and Accomodation Revised.* Studies in History and Philosophy of Science 45 (2014) 54–61

Wray, K. B. (2024), *Kuhn's The structure of scientific revolutions at 60*. Cambridge, Cambridge University Press.

Zahar, E. (1973.1), "Why Did Einstein's Programme Supersede Lorentz's? (I)", in *The British Journal for the Philosophy of Science*, 24.2 (Jun.), pp. 95-123.

Zahar, E. (1973.2), "Why Did Einstein's Programme Supersede Lorentz's? (II)", in *The British Journal for the Philosophy of Science*, 24.3 (Sep.), pp. 223-262.

A note on a Kuhnian-Lakatosian reading of the debate between realism and constructivism in logic

Antonio Piccolomini d'Aragona
University of Tübingen
antonio.piccolomini-daragona@uni-tuebingen.de

1. Introduction

Kuhn's (1962) and Lakatos' (1978) theories have been fertile tools for interpreting, not only the chronological development of scientific ideas, but also and above all the normative structure of science as a whole. Although there is no lack of examples of applications in branches of science other than those where they have proved to be more fruitful, namely, physics and chemistry, it is currently unclear whether they can be used in connection with formal sciences too.

In the case of mathematics, this question is intertwined with that about whether one can licitly speak of revolutions in mathematics, and this is of course because both for Kuhn and for Lakatos, scientific revolutions constitute the keystone for understanding what science is. But if this issue is at least debated in the case of mathematics *in general*, or of branches of it like geometry, analysis, and set-theory, it is instead almost unexplored when referred to that sub-field usually known as *formal*, or *mathematical logic*.

In this somewhat programmatic paper, I aim to provide a first (very much) tentative application of Kuhn's and Lakatos' frameworks to the history of logic. This will not be done, of course, with reference to the history of logic as a whole, but by choosing a specific case-study, i.e. the opposition between realism and constructivism in logic and the foundations of mathematics. The main claim is the following: realism can be looked upon as a Kuhnian paradigm constituted by a semantic level, given by model-theory or similar approaches in the Tarskian tradition, and by a foundational level, given by

Antonio Piccolomini d'Aragona, "A note on a Kuhnian-Lakatosian reading of the debate between realism and constructivism in logic", in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 219-244

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

axiomatic set-theories like ZFC; constructivism is a much more flexible research field, in fact a Lakatosian research programme, which has come in a number of theories, such as Prawitz's semantics (1973, 2015), for the semantic side, and Martin-Löf's intuitionistic type theory (1984) for the foundational level.

Although not the whole history of logic is at issue here, in Section 3 I will nonetheless discuss to some extent the question whether there have been revolutions in logic. This discussion will be in turn framed within an overview of the broader debate about revolutions in mathematics, which I outline in Section 2. In Section 4 I sketch what I take to be the realist and the constructivist pictures, albeit limiting myself to the aspects which seem to me to be more relevant for my purposes. In Section 5 I try to substantiate the view that realism can be looked upon as a Kuhnian paradigm. In Section 6 I do the same for the constructivist field, understood as a Lakatosian research programme. Finally, in the Conclusions I address some potential objections to my claims, and other issues which seem to be raised by the Kuhnian-Lakatosian interpretation I shall be proposing, if one accepts the main lines of it.

2. Kuhn, Lakatos, and revolutions in mathematics: a broad overview

Sciences of different fields develop according to an at least chronological order, consisting of theories which falsify or incorporate previous frameworks, and are falsified or incorporated by the subsequent ones. It is however disputed whether this order is in itself rational, or at least structured. The positivistic idea, dating back to e.g. Comte, that science amounts to a progressive expansion of knowledge and truth, was almost entirely abandoned after Popper (1963) claimed that sciences are "permanent revolutions", given by the interplay of conjectures and refutations.

The revolutionary nature of the history of sciences became prominent with Kuhn's (1962) theory of scientific paradigms. According to Kuhn, while it is true that sciences undergo revolutions, it is instead false that they are *permanently* revolutionary. Rather, proper science is *normal science*, i.e. a puzzle-solving activity in a dominant paradigm. When – for a sufficiently long time – the paradigm fails to explain an anomaly, normal science does become revolutionary. The revolution is accomplished when the revolutionary period yields a new preferred paradigm, which replaces, rather than incorporating, the old one.

For Kuhn, the victory of a new paradigm depends on essentially sociological reasons. It is partly against this that Lakatos (1978) introduced his *research programmes*. The latter do not develop in isolation, but in sequences of theories which may change over time. Changes normally concern only the *protective belt* of a theory, i.e. provisional hypotheses introduced by members of the research programme to adapt their theory to new or unexpected evidence. Renouncing such hypotheses, however, does not alter the *core* of the theory, i.e. the set of fundamental tenets which identify the theory as part of its research programme.

If we accept the research-programmes picture, we can show why and how scientific revolutions can be rationally explained from within science. The "state of health" of the research programmes can be *evaluated* via some parameters, which determine whether sequences of theories are *progressive* or *regressive*, from either an empirical, or a theoretical, or both an empirical and a theoretical viewpoint.

It is much debated whether Kuhn's and Lakatos' theories, as I have roughly outlined them above, can be properly applied to formal sciences. In the case of mathematics, for example, we might argue that Lakatos (1976) himself put forward an approach other than research programmes for explaining how mathematical research grows – on this, see also Moriconi (2022). In a broader perspective, the question may be instead said to boil down to what, with Gillies (1992a), we may call the *Crowe-Dauben opposition*.

According to Crowe (1967, 1992a, 1992b), there is no substantial sense in which we can speak of revolutions in mathematics. Scientific discoveries can be either *transformational*, or *formational*. Only the former are revolutionary, as consisting of radical changes in the structure of a given field. But mathematical discoveries are always formational, namely, they amount to the creation of a new nomenclature, symbolism, or methodology. And these are nothing but surface modifications, which do not alter the historical linearity of mathematics. Dauben (1992a, 1992b) claims on the contrary that revolutions *do* occur in mathematics, and that Crowe employs an excessively strict concept of revolution, according to which revolutions occur only when old theories are *completely* overthrown. But previous configurations may be just relegated to a significantly lesser position, which would be sufficient for marking a discontinuity with the past. For an overview of these and of other important interpretations, the reader may refer to the source book Gillies (1992b), or to the more recent essay Gillies (2023).

More importantly for what is of interest for us here, Oliveri (2006) has looked upon set-theory as what Hallett (1979) called a *mathematical research* *programme*. The core of the programme lies in Cantor's reflections on the mathematical treatment of transfinite or (absolutely) infinite collections. These also provide what Lakatos called the *positive-negative heuristics* of a programme, since they indicate, respectively, which directions the programme should take – e.g. transfinite extension of arithmetic operations – and which must be avoided – e.g. operations on (absolutely) infinite classes. Bueno (2007) has defended the view that Kuhn's incommensurability of theories can be applied to the history of mathematics. According to him, mathematical theories do not develop cumulatively, but via "cuts". The shift from one theory to another implies modifying the *meaning* or the *extension* of given mathematical concepts or predicates. Bueno also brings in some Lakatosian ingredients, as he proposes to read the "cut"-mechanism dialectically: a mathematical conjecture is raised (*thesis*), an alleged counter-example is put forward (*anti-thesis*), a new notion emerges (*synthesis*).

3. Revolutions in logic: Gillies and Kvasz

The works mentioned in Section 2 generally appeal to examples drawn from geometry, analysis or, as we have seen for Oliveri, set-theory. What about that sub-field of mathematics which logic, an up-to-then philosophical discipline, has become starting from the end of the 19th century? A relevant position can be mentioned here, i.e., that of Donald Gillies.

Gillies may be said to be a *discontinuist*, as he believes that one revolution at least has actually occurred in the history of logic, i.e. the *Fregean* revolution (Gillies 1992c). This, however, requires a preliminary adjustment, that is, a distinction between *Franco-British* revolutions – with reference to the British revolution of the 17th century and to the French Revolution of the 18th century – and *Russian* revolutions – with reference to the Soviet Revolution of 1917. In revolutions of the former kind, the old paradigm is not overthrown and, while not being dominant any longer, it retains some degree of importance. In Russian revolutions, the old paradigm is instead set aside forever, and no credit is given to it in the newly established one. Frege's revolution was of a Franco-British kind, as the Fregean paradigm did not reject the old Aristotelian one, but only "embedded" it in a much wider framework. Gillies' interpretation lies in between Kuhn and Lakatos and, as remarked in Gillies (2023), this permits one to have a richer view, not only on the role played by Frege in the revolution that he himself provoked but, additionally, on the contributions that other logicians like Boole, Peano or Russell, gave to the settlement of the new logic.[1]

Gillies' viewpoint may be made compatible with other positions which have instead tended to deny that Frege provoked a true revolution in logic, and which thus conflate Frege's logic with the previous, mostly Aristotelian, tradition. One could for example argue that it is true that *what* logic is expected to do remains roughly the same from Aristotle onward. But, as claimed by Gillies, it is true also that Frege radically changed the conception of *how* logic should fulfil its task. The innovation might be said to stem from something that was absent in Aristotle, i.e. the need of providing foundations, not for science in general, but more specifically for mathematics. It is with this in mind that Frege developed his most important contributions: the abandonment of the Aristotelian dogma that the structure of statements is always in the subject/predicate form, towards a much more comprehensive reading in terms of the function/argument distinction and, as a consequence of this, the introduction of a full-blooded theory of quantification. These can be looked upon as essentially *methodological* innovations. But logic is concerned with language (or with the thought it expresses) so, before being methodological, Frege's novelties can be said to be *linguistic*. Gillies' position may be made compatible with what we may qualify as *continuist* accounts via Kvasz's (2008) theory of the different layers of linguistic innovation in mathematics. The idea would be here, roughly, that Frege's revolution amounted to a Kvaszian linguistic turn out of a content that Frege partially shared with Aristotle. It should be kept in mind, however, that Kvasz's theory is somewhat opposed to Kuhn's picture – albeit partially coping with Lakatos's (1976) one.

4. Realism and Constructivism

After this preliminary overview of (the discussion about) the applicability of Kuhn's or Lakatos' theories to the history of mathematics and logic, let me now turn to the case-study of the opposition between realism and constructivism in logic and the foundations of mathematics.

[1] It should be also remarked that, with respect to the previous 19th century tradition, according to which logic articulated into an *inductive* and a *deductive part*, Frege narrowed the field down to deductive logic only. After this, inductive logic was developed as an independent discipline, where some form or another of probability theory played a major role, while having no place in Frege's deductivist approach. Thus, when below I shall use the expression "Frege's paradigm", one had better read this as meaning "Frege's *deductivist* paradigm". I am indebted to Donald Gillies for this observation. See also footnote 5.

The first remark to be made is that by "realism" I will here mean a number of approaches which jointly operate in both formal semantics and the foundations of mathematics, and which can be said to belong, roughly, to a model-theoretic tradition dating back to Tarski, as regards the semantic side, and to axiomatic set-theory (e.g. ZFC) at the foundational level. *Contra* this, we can identify a constructivist trend which has criticised various features of the realist framework, and which has materialised in a number of theories stemming from the interplay of the intuitionistic and finitist traditions. In the constructivist field too we can distinguish between approaches which are mostly semantic-oriented and approaches which are more interested in foundational issues. I shall here focus on Prawitz's semantics for the semantic side, and on Martin-Löf's intuitionistic type-theory for the foundational side. A certain balance is given thereby to the picture I shall be proposing, as can be seen from the following rough scheme.

	Realism	**Constructivism**
Semantics	Model-theory	Prawitz's semantics
Foundations	Axiomatic set-theory	Martin-Löf's type theory

4.1. An overview of the realist field

When referring to realism, I will first of all make the following historical assumption: a relevant part of the history of modern logic can be split into two macro-stages, i.e.,

- a *foundational stage*, from 1879, publication date of Frege's *Begriffsschrift*, to 1930, when Gödel's incompleteness theorems became known, and
- a *meta-linguistic stage*, from 1930 onward.

In the first stage, the three major foundational schools of logicism, finitism, and intuitionism, aimed at providing mathematics with solid foundations – a need stemming from historical reasons which I cannot deal with here.

Gödel's incompleteness results were a hard blow to all three of these programmes. By showing that no sufficiently powerful (consistent and recursive) system can prove every arithmetical truth, the logicist project of setting up a purely logical calculus capable of reproducing the whole of mathematics was shown to be unattainable. At the same time, by showing that one such truth is the one stating the consistency of arithmetic, Hilbert's idea turned out to be wrong too, as it required precisely a finitary proof of consistency for the ideal-transfinite part of mathematics. And also the intuitionistic identification of truth and provability became problematic, given that Gödel's results are easily read as showing that an unbridgeable gap exists between the general notion of provability and the notion of *formal* derivability – while it is also true, however, that intuitionism was not shown to be *wrong* by Gödel's theorems.

After the Gödelian storm, two (relatively new) theoretical frameworks came to the fore: Tarski's formal semantics was one, the other being axiomatic set-theory, whose most famous and most used account was – and still is today – the one due to Zermelo and Fraenkel – whence the initialism ZFC, where C, as known, indicates the Axiom of Choice.

Tarski's (1956b) semantics starts with a formal definition of the truth predicate. The investigation is tied to a mishmash of findings on the expressibility of the semantic properties of, or in relation to, arbitrary languages, but soon ends up focusing mainly on paradoxes. These are found to arise due to the kind of *semantic closure* that obtains when the semantics of a given language is expressed in the language itself. Tarski then introduces the well-known distinction between *object language*(*s*) and *meta-language*(*s*), so that paradoxes are avoided by requiring the semantic analysis of the object-language not to be carried out in the object-language itself, but in the meta-language.

As Tarski's investigation mainly targets the *formal* languages of the new logic, and the definition of a suitable truth predicate for them, the resulting meta-linguistic semantics will have to be formal as well. This is achieved via *interpretation* functions which map the symbolic constructs onto some base structure(s) – but it should be kept in mind that a precise historical reconstruction of this stage of Tarski's approach risks being biased by the current way of doing model-theory, see e.g. Schiemer & Reck (2013). We are led thereby to the (ancestor of the) notion of *model* of truth or falsity of an (interpreted) formula and, later, to Tarski's own (1956a) adaptation of this machinery to the concept of logical consequence.

The axiomatisation of set-theory is actually much older than Tarski's investigations and Gödel's theorems. If we restrict to ZFC, for example, Zermelo's first attempts at axiomatising the structure and properties of (what would later become the ZFC hierarchy of) sets can be said to articulate within the Hilbertian school and axiomatic method – see Smorynski (2007). Axiomatic approaches like Zermelo's aimed at amending Cantor's original framework, which had been beset by paradoxes similar to Russell's.

However, it was after the collapse of the foundational "schools" that axiomatic set-theory came to play an important, if not privileged role in the foundations of mathematics.[2] Set-theoretic concepts had been around and used – with different meanings and roles, e.g. Frege's "course of values" – by many mathematicians at the end of 19th century but, even in the framework of Hilbert's Programme, set-theory was looked upon as just a *part* of mathematics. And precisely like the whole of mathematics, set-theory too was understood as standing in need of foundations, not as the framework where mathematics could be provided a firm foundation with.

It was in this connection that axiomatic set-theory ended up being inextricably intertwined with Tarski's account, leading to what would eventually become model-theory. Tarski's structures simply became sets equipped with functions and relations, out of some axiomatised (mostly ZFC) universe. Model-theory and set-theory started overlapping significantly and, while not identifying with each other, they exchanged such a number of notions and results that, for the large part of today's logicians, it would be impossible to be acquainted with the former without having at least some familiarity with the latter – where it is remarkable that the inverse does not hold, meaning that the *foundational* role is played by set-theory *towards* model-theory. Clearly, this does not mean that such a connection *had* to occur, only that it *de facto* happened. More importantly, the notions and results that set-theory and model-theory exchanged were generally referred to classical logic and classical mathematics. To put it roughly, they presupposed some background *bivalent* notion, e.g. a bivalent notion of truth to the effect that every truth-bearer is either true or false, whence excluded middle holds.

4.2. An overview of the constructivist field

This realist picture was not immune from criticisms. Philosophers, logicians and mathematicians concerned with *epistemic* issues argued that the logical laws and the foundations of mathematics should not be detached by knowledge and computational control over the foundational principles.

[2] This is not to say, however, that axiomatic set-theory was not acknowledged as capable of encompassing the whole of mathematics. But in Zermelo's times, the same was for example true also of Russell's type theory, see Lolli (2011). Attempts at considering axiomatic set-theories as a foundation of the whole of mathematics were also there much before Gödel proved his theorems but, besides the parallel and somewhat contrary idea that such theories should be *in turn* provided a (proof-theoretic) foundation with, set-theoretic foundationalism underwent many objections, the best known example being perhaps Skolem (1922). On this see also Kahle (2015).

Constructivists usually develop formal semantics and foundational approaches to mathematics where the notion of (bivalent) truth is replaced by the notion of proof, and where sets are postulated to belong to universes over which a kind of "computational" control is guaranteed. These aspects are then connected via the idea that the notion of set and the notion of formula are indistinguishable, since a formula amounts to the class of its proofs, and a class determines univocally the proposition which says that that class is inhabited, thus being the space of the proofs of this formula, so both the proposition and the class are *types*. This is called the *formulas-as-types conception* which in turn, via Howard (1980), is the basis of the *Curry-Howard isomorphism*.

Prawitz and Martin-Löf are prominent figures here.[3] Besides the Curry-Howard isomorphism, we can mention a number of other sources they share. One of these is BHK-semantics, the intuitionistic proof-based meaning explanation of the logical constants – as for example found in Troelstra & van Dalen (1988). Next to this, we have Gentzen's *proof theory*, qualified as *general* by Prawitz (1973), to distinguish it from Hilbert's *reductive* proof theory – others have shared this interpretation, see Cellucci (1978), Moriconi (1988) or Schroeder-Heister (2006). While Hilbert aimed at showing that results in certain (ideal) parts of mathematics reduced to results in other (real) parts of mathematics, Gentzen aimed instead at studying proofs *as such*, namely, at investigating their structural and, so to say, geometrical properties – although Gentzen's project is still best framed within Hilbert's broader Programme, see e.g. von Plato (2012). Gentzen's own (1935) crucial proof-theoretic results mainly pertained to one of the two kinds of calculi that he had invented, that is, Sequent Calculus, and they were later on expanded by Prawitz (1965) in his doctoral dissertation as *normalisation theorems* for (various systems of) Natural Deduction – the other calculus introduced by Gentzen.

[3] Let me specify that the choice to limit the analysis of the constructivist field to Prawitz's proof-theoretic semantics and to Martin-Löf's type theory is surely arbitrary, but not without reasons. It is arbitrary in that, as said, constructivism is exemplified by a number of theories which, although sharing some principles, may differ quite substantially from each other. Applications of the interpretive line that I am proposing in this paper to constructivist approaches other than Prawitz's or Martin-Löf's might be carried out in future works. The choice is not without reasons either, for (to my mind at least) Prawitz's semantics and Martin-Löf's type theory are, contrarily to others, two *strictly intertwined* approaches (not only in the conceptual, but also in the historical sense) of which one is, as said, more semantically oriented, while the other one is more foundational in spirit. In this way, we can attain that balance with the realist counter-part, which I have discussed in beginning of this section.

In the 1970s, Prawitz (1971, 1973) began to develop a semantics out of his normalisation results. Prawitz's semantics, called *proof-theoretic semantics* by Schroeder-Heister (1991), is in many ways opposed to Tarski's approach. In line with the constructivist desiderata, the central notion is not that of truth, but that of *valid argument*, where arguments are chains of *arbitrary inferences*. Arguments are labelled by *reductions* turning input arguments from assumptions S to conclusion A into output arguments from assumptions S* ⊆ S to conclusion A. Output arguments are expected to enjoy some relevant semantic features. In particular, since the introductions in Gentzen's Natural Deduction are looked upon – as Gentzen (1935) himself suggested – as definitions of the meaning of the logical constants they concern, whence they are valid by default, all inferences other than the introductions must be justifiable by showing that they are harmonic relative to how meaning is determined by the introductions. Examples of such reductions are precisely those used by Prawitz himself for showing how redundant steps could be eliminated from Natural Deduction derivations, e.g., in the case of a conjunction-detour,

$$\dfrac{\dfrac{\begin{matrix}\Delta_1 & \Delta_2 \\ A_1 & A_2\end{matrix}}{A_1 \& A_2}}{A_i} \quad \Longrightarrow \quad \begin{matrix}\Delta_i \\ A_i\end{matrix}$$

($i = 1, 2$). Thus, an argument can be said to be *valid* when, roughly, it reduces by iterate applications of the reductions which it is labelled with to an argument ending by introduction whose immediate sub-arguments are valid.

Martin-Löf's type theory has come in many versions. Here I shall refer mainly to the (1984) one. It can be understood as a broad framework for encompassing mathematical knowledge and, thereby, for providing it with a foundation. Logic is conceived of as a sort of by-product of this foundation, although not in the sense that Martin-Löf subscribes to some Hilbertian standpoint – but some have spoken in this connection of a *constructive Hilbertian programme*, e.g. Rathjen (2005). The theory involves a family of increasingly powerful systems, which can be generated following some basic principles. The latter stem from constructivist tenets about what it means to be justified in making judgements like "A is a set", "*x* is an element of A", "A and B are equal sets" and "*x* and *y* are equal elements of A". The set-judgements are explained by giving conditions for forming canonical elements, whereas for the element-judgements one requires that, for it to be an element of a set, *x* must evaluate – i.e. compute – to a canonical element of the set. This yields four basic rule-forms: formation rules (set-formation), introduction rules (canonical elements), elimination rules (operators for evaluation) and equality rules (equations which define operators). A distinctive trait of Martin-Löf's theory is that it allows for *dependent types* or *objects*, namely, for the formation of sets and elements where free variables may occur, and on which one can make assertions given hypothetical knowledge about the status of those variables. E.g. "B(*x*) is a set, under the assumption that *x* is an element of A", "*a*(*x*) is an element of B, provided *x* is an element of A", or combinations of these. One may then introduce rules for logical operators, like Cartesian product of family of sets, disjoint union etc., or add new sets, both of an atomic kind and of a more complex kind.

In line with the Curry-Howard isomorphism, sets in type theory are understood as types, and are equated with propositions which suitable proof-objects are elements of. Types and elements of types must be generated in a constructive manner, to the effect that, if compared with "realist" set-theory, a type-theorist would require a much stricter control over the "ontology" of the universe underlying its theory – see e.g. Klev (2019). Perhaps more importantly, Martin-Löf renounces a distinction that, since Tarski's treatment of semantic paradoxes and Gödel's incompleteness theorems, has become a standard feature of logical analysis, namely, the separation of object-language and meta-language, or of syntax/deduction and semantics. In type theory, the language is "built from within the system", in such a way that the validity of the rules employed be immediately evident. The acknowledgement of this validity is then conceived of as, say, contemporary to the process of deriving judgements in the systems, the latter being, contrarily to the ontological nature of proof-objects for propositions in types, a purely epistemic activity. This is similar to what happens in Frege and Russell approaches, where systems come with an intended meaning, *contra* Hilbert's idea of re-interpreting axioms over different models – see Sundholm (2001) whereas, for the Hilbertian proto-notion of model see Eder & Schiemer (2018) and Schiemer & Giovannini (2024).

5. Model-theory and set-theory as a Kuhnian realist paradigm

In this section, I want to provide some programmatic hints at how to substantiate the claim that model-theory and axiomatised set-theory constitute a *realist paradigm* in Kuhn's sense. The claim will be split into two sub-claims, i.e.:

1. model-theory and axiomatised set-theory have been normally carried out under a (possibly unacknowledged) philosophical perspective which may be qualified as realist, and
2. they have come to form a Kuhnian paradigm.

Both 1 and 2 will be (tentatively) justified in a mostly *historical*, rather than *conceptual* way. While this may be obvious for 2, it is less so for 1. That model-theory and set-theory have *happened* to be used or developed from a normally realist standpoint, does not also imply that they *had* to be used or developed thus.

Any attempt at providing a uniform and rigorous characterisation of what realism is, would be of course far beyond the scope of this paper. For my purposes, it will suffice to conceive of realism as given by two interrelated principles. The first is the already mentioned principle of *bivalent truth*, according to which every truth-bearer is determinately either true or false. The second principle is the idea that in logic and the foundations of mathematics we can freely refer to "facts", objects and structures which are independent of our ability to effectively construct, know, or ascertain them.

It seems to me to be unquestionable that model-theory and axiomatic set-theories like ZFC have been historically understood as respecting the principle of bivalent truth. Model-theory is normally assumed to be a semantics where bivalence holds, which in turn is mirrored by the fact that models are precisely (structured) sets from an axiom system whose logic also validates bivalence. And even without any overlap of models and sets, Tarski himself seems to understand the universe underlying his early semantic investigations as one where bivalence holds, so to say, by default, thus providing a (partly circular) justification of the unrestricted validity of the law of excluded middle. As for the unknowability or non-constructibility of "facts", objects and structures that logical and foundational investigations refer to, these may be said to stem from the domain(s) which the set-axioms are to be interpreted onto. The ZFC-hierarchy, for example, involves properties or "entities" which lie well beyond any effective possibility of building them up in a constructive manner, and which thus exist independently of any in-principle capability of coming into possession of them.

Let me now turn to point 2. In what sense model-theory and set-theory can be said to constitute a Kuhnian paradigm? If we accept Gillies' view that

Frege determined a logical revolution, we should also conclude that model-theory and set-theory relate to some Fregean picture. Before addressing the issue whether model-theory and set-theory can be *actually* understood as a paradigm, we must first ask how, *if* so understood, model-theory and set-theory interact with a logical picture stemming from Frege. Here, we seem to have only three options available: first, the logical and foundational analyses that the intertwinement of model-theory and set-theory produced are a *sub-paradigm* of a Fregean paradigm; second, the intertwinement is in fact an entirely *new* paradigm; or third, we are in the presence of an *evolution* of the Fregean framework, of the kind of those stemming from pre-revolutionary periods when an approach is capable of self-adjusting against some serious anomaly.

I think that the second option can be discarded outright. This is not only because model-theory and set-theory are fundamentally indebted to the *modus operandi* inaugurated by Frege or by those walking in his footsteps (Hilbert and his school included) – just think for example of the languages which model-theory and set-theory normally employ, or of the proof-systems over such languages.[4] The main point is that, if model-theory and set-theory are to be understood as a new paradigm, they should have become so via a replacement of some *previous* paradigm after a crisis and a subsequent revolutionary period. And one could very hardly maintain that model-theory

[4] Concerning this, it must be remarked that speaking of a *Fregean picture* in logic and the foundations of mathematics may be too restrictive. Both model-theory and set-theory are influenced by two *Hilbertian* ideas which are not to be found in Frege – at least not *prima facie*. The idea that formal languages can be taken as sets of meaningless strings to be interpreted onto different domains, which is the core idea of model-theory, is often said to stem from Hilbert's *Grundlagen der Geometrie*, i.e. from the idea that the language of a given axiom-system can be attributed a meaning by associating linguistic components to suitable entities and properties from different (mostly numerical) fields. Hilbert's original aim was that of proving independence or consistency results for (groups of) geometrical axioms – see, e.g., Eder & Schiemer (2018) – but, later, the strategy was implemented into Hilbert's Programme, as part of the idea that formal languages and proofs should be (meta-mathematically) studied as mathematical objects in themselves. Axiomatic approaches to set-theory, instead, first developed by Zermelo in a fully systematic way, took place within the axiomatic tradition inaugurated by Hilbert and his school, which Zermelo can be said to be an active and prominent member of – see e.g. Moriconi (1976, 1988) and Sieg (2013). Thus, Gillies' thesis about the existence of a Fregean revolution in logic may have to be refined by speaking of a Fregean-Hilbertian revolution rather, and accordingly of a corresponding Fregean-Hilbertian picture. Perhaps, the expression "classical logical approach" would fit better, with the proviso that "classical" should be not understood as referring to classical logic.

and set-theory replaced Frege's approach, the latter being the only reasonable candidate for playing such a role.

We are hence left with the first and the third option. Can the semantic-foundational framework of model-theory plus set-theory be understood as a proper part of Frege's legacy? This also seems to me to be wrong. For, there are at least two features which model-theory and set-theory enjoy, and which can be said *not* to be part of Frege's account.

The first, mainly in the realm of model-theory, concerns the clear separation between the logical language and its semantics. The first logical languages – for example, those of Frege and Russell – were meaningful, in the sense that they were understood as coming with an intended interpretation. The idea that logical languages are mere "algebraic structures", standing in need of, say, external interpretation, appears only later. According to many, the roots of it can be traced back to Hilbert's *Grundlagen der Geometrie*, but it was only with Tarski's paper on the concept of truth in formalised languages (1956b) that the distinction of syntax and semantics became, so to say, indispensable. It was perceived, not merely as a possible or fruitful distinction, but as "real" and deep one, having to do with crucial aspects and constituting the *necessary* condition of any logical investigation. As argued by, e.g., Sundholm in various papers (2001, 2009, 2019), such a "meta-linguistic dogma" came as an answer to the crisis provoked by the discovery of paradoxes like Russell's, and by the effects that Gödel's incompleteness results had on foundationalism.

The same reasons may be said to be behind the second feature, which instead mostly pertains to ZFC. This is the idea that mathematics should not be given foundations through broad philosophical principles inspiring some foundational project, but on a *specific* axiomatic theory, i.e., an axiomatic theory for sets. In turn, this is not understood in a *reductionist way* – say, from mathematics to logic, or from transfinite to finitary theories – but *assuming* that the foundations is achieved as soon as the reference theory has been suitably axiomatised.

If one accepts that this "theoretical surplus" is part and parcel of the current mainstream way of carrying out logical investigations, then one should also conclude that model-theory and set-theory are best seen as an *evolution* of Frege's approach, i.e., as a self-adjustment of it after some anomalies that it was not able to get rid of *per se*.

Let us now finally turn to the issue whether model-theory and set-theory can be understood as giving rise to a paradigm in Kuhn's sense. My discussion of this claim will be twofold: historical and "statistical". The historical reasons that led to the spread of model-theory and set-theory are of course too many to be explored, but an essential list should in my opinion include the following at least:

- completeness proof of first-order logic as provided by Gödel in 1929. Gödel employs (what can be seen as) a notion of model. Also, in subsequent conversations with Wang, he says that his discovery was made possible by leaving out the constructive *desiderata* which prevented Skolem to obtain the result – see e.g. Kennedy (2020) and Wang (1996);
- the satisfactory treatment of the semantic paradoxes in the Tarskian approach, and of the set-theoretic ones in the Zermelo-Fraenkel axiomatisation;
- the clear-cut distinction between derivability in formal systems and "truth", which seems to cope perfectly with a reading of Gödel's incompleteness results as showing that the concept of provability and the concept of derivability in a calculus cannot coincide.

As for the "statistical" reasons, besides the trivial observation that model-theoretic and set-theoretic approaches are (quantitatively) dominant today, the following facts seem to me to hold:

- some seemingly open problems are completely left aside, or else dealt with under a suitable adaptation to the tools and principles already at play in the background framework. A case of this are *epistemic considerations*, for example, the requirement that the modality involved in (logical) consequence, when saying that it amounts to *necessary* truth-preservation (under variations of the meaning of the non-logical terminology), should be understood in terms of epistemic compulsion. This forces a distinction between the notion of (logical) consequence and the notion of (logically) valid inference. Now, these issues are rarely addressed (or even ascertained) in standard semantic approaches or, if they are, this happens in an "extensional" way, say in terms of epistemic operators, which is unsatisfactory, e.g., to most constructivist logicians;
- we have an almost complete fulfilment of Kuhn's *textbook criterion*. Logical textbook use to provide a model-theoretic semantics, or variants of it. Also, the first sections of textbooks in practically *any* field of mathematics (say, analysis, algebra, geometry, and so on) start with a (not always axiomatic) presentation of the basic notions of set-

theory. Almost none of these textbooks mentions the existence of alternative approaches;

- as concerns ZFC, it has been extensively used in the 1950s, 1960s, and part of the 1970s, as a basis for reforms in mathematical education in most countries from all over the world – a very useful source here is in my opinion Pellerey (1989).

6. Constructivism as a Lakatosian research programme

Before addressing the issue about whether and in what sense Prawitz's proof-theoretic semantics and Martin-Löf's type theory constitute a constructivist Lakatosian research programme, let me first outline a notion of *logical research programme*, partly along the lines of Hallett's (1979) notion of *mathematical research programme*. This is not meant to be an exhaustive discussion, though, and a more refined treatment might be carried out in future works. However, I will provide examples that, hopefully, clarify the concepts I shall be introducing.

A logical research programme is given by a (possibly ramified) sequence of logical approaches, linked to each other by a number of principles, methods, results, and open problems. The *core* of the sequence amounts to a number of *informal desiderata* which the logical investigations are modelled on, whereas the *protective belt* is given by a number of more or less rigorous (formal and informal) hints at how the link between the logical analysis and the informal desiderata might be attained. The sequence will of course not be theoretically or empirically progressive in the same sense as one of physical or chemical theories would be. There are no theoretical predictions to be made, nor empirical outcomes to be confirmed or refuted in the outer world. These notions are, so to say, fully internal to the sequence. The latter can be said to be *theoretically progressive* when the basic assumptions of the analyses it amounts to provide a "glimpse" into relevant conceptual frameworks concerning crucial notions of the logical field, or in that part of this field that the sequence is dealing with. If we give a Lakatosian look at the foundational programmes, for example, some instances of this can be seemingly found: Frege's innovative proposal that the concept of number could be defined in purely logical terms; Hilbert's creation of proof-theory and of proof-theoretic techniques; Brouwer's "vow" to bring a new mathematics about. The sequence can be said to be *empirically progressive* when the set of "concrete" results it achieves seemingly shows the fruitfulness of the approaches it includes. These are theorems, or else definitions of previously vague ideas. Again with reference to foundationalism, we may mention: Frege's *actual* logical definition of the concept of number, and the role that such definition played both in his (inconsistent) theory, and in Russell's encompassing type theory; Hilbert's partial results towards a proof of consistency of arithmetic and their by-products (like the ε-theorems), which were used in proof-theory even after the abandonment of Hilbert's programme; Brouwer's definitions of non-trivial notions such as those of *choice sequence* or *bar induction*, or the proof of such results as the *fan theorem*.

The core tenets guiding the research of a logical research programme "transcend", in a sense, the actual development of the research itself, as they, say, shape the direction which this development is expected to take. This is what happened for Frege's claim that mathematics is overall reducible to logic, or for Hilbert's stance that transfinite, ideal mathematics should be shown to be conservative or consistent over finitary, real mathematics, or finally for the broad intuitionistic philosophy which animated Brouwer's research. The resistance to question principles in the core can be clearly seen in these cases. E.g., the discovery of Russell's paradox did not lead to renounce logicism, but rather to modify the way the foundational programme had to be carried out (i.e. part of the protective belt). Likewise, Gödel's incompleteness results did not lead Hilbert or his fellows to renounce finitism, but to change their point of view on the extent of finitary mathematics (even to introduce a kind of ω-rule). One might even claim that a modified Hilbert's programme is still in place – see e.g. Sieg (2013).

It should be clear from the above that I take the *foundational part*(*s*) of Frege's paradigm to form a (ramified) research programme. The above-mentioned anomalies led to split this part from the one which could be kept in the re-adaptation of the paradigm via model-theory and axiomatic set-theory, i.e., what we may call the *methodological* component (including for example the development of a certain logical grammar, or of certain axiom systems, or the requirement that certain results *had* to be proved, like consistency, categoricity, completeness of various types, etc.). The foundational part was abandoned in the re-adaptation (or else it was re-framed as foundations onto axiomatic theories for sets). However, pieces of the foundational programmes survived, and branches of the sequence which they were part of flowed into approaches alternative to the semantic-foundational paradigm of model-theory and set-theory. Constructivism can be understood as one such approach since, so it seems to me, it can be understood as given by a kind of mix of some intuitionistic and some finitist tenets, among which:

- the intuitionistic principle that meaning should be explained via proof-conditions, eventually leading to the rejection of bivalence and, hence, of excluded middle, and of the non-classical use of the existential quantifier, and
- the finitist claim that an upper bound must be required for the computational complexity of one's proof-methods.

These aspects are to be found both in Prawitz and in Martin-Löf. The idea that meaning is explained in terms of proof-conditions becomes in Prawitz the idea that meaning is given by some (privileged) rules. In Martin-Löf, the same holds for propositional proof-objects, but here we also have the idea that meaning of judgements is given by stating conditions under which judgements can be correctly made. The requirement of computational control on proof-methods becomes in Prawitz the idea that reductions have to be effective functions, so that reduction sequences of proof-structures be actually computable. In Martin-Löf, introduction of, e.g., higher types (say, *universes*) is allowed only via some form of reflection over families of types of lower level (although this may be said to hold for the introduction of *any* type) – see Rathjen (2005) and Klev (2019).

Building on this shared background, Prawitz's proof-theoretic semantics and Martin-Löf's type theory have developed throughout the years following a number of programmatic and quite flexible lines. Also, they have often intertwined: Martin-Löf has looked at Prawitz's normalisation results when putting reducibility constraints on propositional proof-objects, or for elimination and equality rules over given types; Prawitz has understood Martin-Löf's type theory as a framework where a number of fundamental issues could be fruitfully used for his own semantics, the last example being his idea of explaining inferential validity via epistemic grounds which can be understood as proof-objects in type theory – based on the common endorsement of the *formulas-as-types* conception.

There are of course a number of divergences, as there are divergences internal to the Prawitzian and Martin-Löfian fields. The latter gave often rise to approaches which, while still Prawitzian and Martin-Löfian in nature, are *substantially* different from each other. But this is something one may after all expect from the kind of "fluidity" that a Lakatosian research programme should show, and which seems not to happen (at least not to the same extent) in the well-established model-theoretic or set-theoretic tradition.

If we accept to consider the articulation of Prawitz proof-theoretic semantics plus Martin-Löf's type theory as a logical research programme in Lakatos' sense as outlined above, we cannot fail to observe, however, that those two approaches have rather different aims. While Prawitz's semantics, in a meta-theoretic perspective, aims at explaining constructively the meaning of a logical language, and at providing on this basis a constructive notion of (logical) consequence, with respect to which soundness and completeness results are proved, Martin-Löf has no meta-language, and his project can be understood as a foundational one which, in Sundholm's (1994: 37) words, amounts to an "effort towards a realisation of a constructivist theory of meaning for an […] interpreted language serving the needs of pure mathematics". Hence, also the kind of results that one expects the overall programme to yield to prove its fruitfulness, are not on par. To mention just few examples: people working in a properly Prawitzian field, tend to address typically semantic issues, such as logicality, completeness, and extendibility of the approach to natural language – see e.g. Francez (2015); as for type theory, it gave instead rise to the much more foundational project of *Homotopy Type Theory* (2013) – although examples of applications to the natural language can be significantly found in this case too, see Ranta (1994).

This notwithstanding, it seems to me to be undoubted that Prawitz's semantics and Martin-Löf's type theory share a number of principles, which may be said to form, or at least be part of the core of a constructivist research programme, and whose precise development, in the protective belt, is the aim of the formal enterprises they amount to. These principles concern *epistemic issues*, which lie at the basis of the common criticism of approaches based on model-theory and set-theory, in turn perceived as realist in the broad sense. They also dictate, influencing how the machinery in the belt is to be devised, the positive and negative heuristics of the programme.

So far I have said nothing about how the heuristics of a logical research programme should be understood. Dealing with this issue in deep would require much more details than what this paper – which, as said, is to be understood as programmatic – is expected to do. Thus, I will discuss these topics in future works, but here I want to provide one example of negative heuristics relative to Prawitz's semantics. As said, it is part of Prawitz's project that some computational constraints are put on the kind of reductions via which the validity of given proof-structures can be established. This depends on the epistemic concerns from the core: since the overall approach must cope with such ideas as that logical consequence is not truth-preservation, but a modal epistemic link between truth-bearers, and since this modal link is to be given in terms of proofs, the semantic formal counterparts of proofs (i.e. valid arguments) must be recognisable as such, in the sense that they must reduce to semantically privileged forms in a *decidable* way. Now, this decidability cannot be normally achieved, not even when reductions are taken to be effective functions. One reaction to this could be to claim that there can be no such thing as developing a proof-based semantics where proofs are the kind of objects that Prawitz would like them to be – essentially, objects which are formal in nature, but not reducible to derivations in a system. But this would mean throwing away one of the most crucial principles in the core (together with additional features which are proper to Prawitz's project). And in fact, *this has not been* Prawitz's reaction. Rather than touching the core, Prawitz has preferred to seek changes in the protective belt, by requiring additional constraints to be put on reductions so as to limit their potential complexity – see e.g. Prawitz (2019a, 2019b) – or even by renouncing the idea that a non-circular definition could be given of the intertwined notions of valid inference and proof – see Prawitz (2024).

7. Conclusions

Due to the programmatic nature of this paper, what I have been saying admittedly leaves many questions unanswered. I would like to suggest potential further developments of the line of research I have been sketching, together with potential objections which might be raised against it, and with potential ways of meeting these objections.

First, the notion of logical research programme in Lakatos' sense should be characterised more precisely. I have said few things about how a sequence of logical approaches could be understood, as well as about what the core and the protective belt of such a sequence might be taken to be. I have said even less about the notions of positive and negative heuristics of a logical research programme. These issues can be dealt with further in future works.

Concerning the previous point, one might reasonably claim that the development of model-theory and set-theory can be in turn read in Lakatosian terms, hence as constituting a *realist* logical research programme (*contra* the idea, argued for in this paper, that they instead form a realist logical paradigm). This seems to be for example the position of Oliveri (2006) and, partly, also of Bueno (2007). To my mind, such a reading is, not only reasonable, but even correct. For nothing impedes that, either as separate mathematical theories or as joint logical frameworks, model-theory and set-theory have been for a number of years a logical research programme and that then, due to historical reasons and achieved results, they have come to play the role of dominant paradigm.

This in turn requires a more refined characterisation of the relation between the foundational period in logic, on the one hand, and what I have called the realist logical paradigm and the constructivist research programme, on the other. As partly anticipated above, it seems to me that the pre-Gödelian investigations are animated by two main attitudes: a *foundational* one and a *methodological* one. The former was *discarded* in the realist paradigm, and replaced by the idea that foundation required a *specific* axiomatic theory; it was instead *modified* in the constructive research programme, by re-reading some peripheral principles having to do with meaning and computational control. The *methodological* part mainly concerns the way one carries the logical analysis out (logical languages, axiom systems, etc.), and it seems to be common to both approaches (although Martin-Löf's notion of judgement is also partly connected to a *pre*-Fregean tradition).

It is in this respect crucial to assess, in either Kuhnian or Lakatosian terms, the picture stemming from Frege's work – or better, from Frege's *and* Hilbert's work, see footnote 4. In particular, can one identify something like a Fregean *paradigm*, or had one better speak of a Fregean *research programme* (or none of these)? More in general, is the foundational period best understood as a *paradigmatic*, or a *research-programmatic* one? Does this hold for the period *as a whole*, or must the assessment be adapted to one or the other of the foundational schools? Finally, is the answer the same for both what I have called the foundational and the methodological part of the pre-Gödelian approaches, or can one of these be said to be Kuhnian, the other being Lakatosian?

I personally and currently believe that the foundational period is best looked upon along the lines of Lakatos' theory, but I will not articulate this standpoint here since, as the reader might have easily realised, it makes it very difficult to describe how constructivism, seen as a research programme on its own, stems from the previous approaches: given that, as said above, the foundational component of the foundational framework is *rethought* by constructivism, it is likely that the core of the latter contains new principles (or discards, or reorganises part of the old ones). The question is therefore whether the constructivist research programme is an evolution of the foundantionalist programme, or an entirely *new* programme. Answering these questions is far beyond the intentions of my paper, and it would require much more space than what is allowed for here. Thus, I will (try to) address them in future works.

Similar questions concern the respective relations between model-theory and set-theory, on the one hand, and the Fregean tradition on the other. As said, I do believe that it is fair to conceive of model-theory *plus* set-theory as research programmes which, *later on*, gave rise to a unified realist paradigm. A Lakatosian reading of the foundational schools seems to be less problematic here, since one could say that model-theory and set-theory constituted a new research programme, by replacing old principles in the (foundational part of the) previous core with some new stances. But this also implies that one must find a way to articulate a framework where Kuhn and Lakatos can *co-exist*.

In fact, a main objection to the approach I am proposing here is that one cannot have such a mixed picture. In analyses of this type, one should go either totally Kuhnian, or totally Lakatosian. This is true. But it is also true that one may conceive a mixed approach where one does not really have a *co-existence* of a Kuhnian paradigm and a Lakatosian research programme, but a co-existence of two approaches where Kuhnian paradigms are informed by Lakatos' research programmes, and vice versa. Such an approach has been suggested by Gillies (1992b, 2023) and, as for logic, it may benefit from Kvasz's (2008) theory of linguistic changes in mathematics.

This topic will be addressed in future works too. The rough lines of it, of which I can give only a quick sketch here, are however to the effect that the overall approach becomes basically *Lakatosian*. On top of it, one then introduces some *strength* parameters, for distinguishing between a (more or less) *rigid* and a (more or less) *flexible* research programme. A rigid research programme is one where certain sociological conditions, such as those satisfied by a Kuhnian paradigm, are fully met, so that, say, the core is relatively big (it is harder to make it collapse), while the protective belt is relatively small (it is easier to spot counter-examples). In a flexible research programme, instead, we have the inverse situation: the core is relatively smaller, while the protective belt is relatively bigger, and the sociological factors identifying a Kuhnian paradigm (or most of them) fail. The idea is, all in all, to complement the *internal* (epistemological) reading of a series of scientific theories, provided by Lakatos, by means of *external* parameters, drawn from Kuhn, and to provide thereby a *Kuhnian evaluation* of the rigidity/flexibility of a *Lakatosian* research programme.

Acknowledgements

Work on this paper was financially supported by the grant PI 1965/1-1 for the DFG project *Revolutions and paradigms in logic. The case of proof-theoretic semantics*.

References

Bueno, C. (2007), "Incommensurability in mathematics", in B. van Kerkhove & J.P. van Bendegem (eds), *Perspectives on mathematical practices*, Logic, epistemology and the unity of science, vol. 5, Springer, pp. 83-105.

Cellucci, C. (1978), *Teoria della dimostrazione: normalizzazioni e assegnazioni di numeri ordinali*, Bollati Boringhieri.

Crowe, M. (1967), "Science a century ago", in F. J. Crosson (ed), *Science and contemporary society*, University of Notre Dame Press, pp. 105-126.

Crowe, M. (1992a), "Afterword (1992): a revolution in the historiography of mathematics?", in D. Gillies (ed), *Revolutions in mathematics*, Clarendon Press, pp. 306-316.

Crowe, M. (1992b), "Ten "laws" concerning patterns of change in the history of mathematics", in D. Gillies (ed), *Revolutions in mathematics*, Clarendon Press, pp. 15-20.

Dauben, J. (1992a), "Appendix (1992): revolutions revisited", in D. Gillies (ed), *Revolutions in mathematics*, Clarendon Press, pp. 72-82.

Dauben, J. (1992b), "Conceptual revolutions and the history of mathematics: two studies in the growth of knowledge", in D. Gillies, *Revolutions in mathematics*, Clarendon Press, pp. 49-71.

Eder, G. & Schiemer, G. (2018), "Hilbert, duality, and the geometrical roots of model theory", in *The review of symbolic logic*, 11(1), pp. 48-86.

Francez, N. (2015), *Proof-theoretic semantics*, College Publications.

Gentzen, G. (1935), "Untersuchungen über das logische schließen I II", in *Mathematische Zeitschrift*, 39:176-210, 405-431.

Gentzen, G. (1936), "Die Widerspruchsfreiheit der reinen Zahlentheorie", in *Mathematische Annalen*, 112, pp. 494-565.

Gillies, D. (1992a), "Introduction", in D. Gillies (ed), *Revolutions in mathematics*, Clarendon Press, pp. 1-14.

Gillies, D. (1992b), *Revolutions in mathematics*, Clarendon Press.

Gillies, D. (1992c), "The Fregean revolution in logic", in D. Gillies (ed), *Revolutions in mathematics*, Clarendon Press, pp. 265-305.

Gillies, D. (2023), *Lakatos and the historical approach to philosophy of mathematics*, Cambridge University Press.

Kahle, R. (2015), "Gentzen's consistency proof in context", in R. Kahle & M. Rathjen (eds), *Gentzen's centenary*, Springer, pp. 3-24.

Kennedy, J. (2020), *Gödel, Tarski and the lure of natural language: logical entanglement and formalism freeness*, Cambridge University Press.

Klev, A. (2019), "A comparison of type theory with set theory", in S. Centrone, D. Kant & D. Sarikaya (eds), *Reflections on the foundations of mathematics*, Springer, pp. 271-292.

Kuhn, T. (1962), *The structure of scientific revolutions*, University of Chicago Press.

Kvasz, L. (2008), *Patterns of change. Linguistic innovations in the development of classical mathematics*, Birkhäuser.

Hallett, M. (1979), "Towards a theory of mathematical research programmes I II", in B*ritish journal for the philosophy of science*, 30, pp. 1-25, 135-159.

Howard, W. (1980), "The formulas-as-types notion of construction", in J.R. Hindley & J.P. Seldin (eds), *To H.B. Curry: essays on combinatory logic, lambda-calculus and formalism*, Academic Press, pp. 479-490.

Lakatos, I. (1976), *Proofs and refutations. The logic of mathematical discovery*, Cambridge University Press.

Lakatos, I. (1978), *The methodology of scientific research programmes. Philosophical papers vol. I*, J. Worall & G. Currie (eds), Cambridge University Press.

Lolli, G. (2011), *La guerra dei trent'anni. Da Hilbert a Gödel*, ETS.

Martin-Löf, P. (1971), "Hauptsatz for the intuitionistic theory of iterated inductive definitions", in J. E. Fenstad (ed), *Proceedings of the second Scandinavian logic symposium*, Elsevier, pp. 179-216.

Martin-Löf, P. (1984), *Intuitionistic type theory*, Bibliopolis.

Martin-Löf, P. (1985), "On the meaning of the logical constants and the justification of the logical laws", in *Atti degli incontri di logica matematica*, vol. 2, Università di Siena, pp. 203-281.

Moriconi, E. (1976), *Alle origini della teoria della dimostrazione di Hilbert*, in *Il Pensiero*, XXI, 1-3, pp. 149-175.

Moriconi, E. (1988), *La teoria della dimostrazione di Hilbert*, Bibliopolis.

Moriconi, E. (2022), "On Lakatos's decomposition of the notion of proof", in G. Oliveri, C. Ternullo & S. Boscolo (eds), *Objects, structures and logics*, Boston Studies in the Philosophy and History of Science, 339, Springer, pp. 331-348.

Oliveri, G. (2006), "Mathematics as a quasi-empirical science", in *Foundations of science*, 11, pp. 41-79.
Pellerey, M. (1989), *Oltre gli insiemi: nascita, crescita e crisi dell'insiemistica*, Tecnodid.
Popper, K. (1963), *Conjectures and refutations. The growth of scientific knowledge*, Routledge.
Prawitz, D. (1965), *Natural deduction. A proof-theoretical study*, Almqvist & Wiskell.
Prawitz, D. (1973), "Towards a foundation of a general proof theory", in P. Suppes, L. Henkin, A. Joja & G. C. Mosil (eds)*, Proceedings of the Fourth International Congress for Logic, Methodology and Philosophy of Science, Bucarest 1971*, Studies in Logic and the Foundations of Mathematics, vol. 74, Elsevier, pp. 225-250.
Prawitz, D. (2015), "Explaining deductive inference", in H. Wansing (ed. by), *Dag Prawitz on proof and meaning*, Outstanding Contributions to Logic, Springer, pp. 65-100.
Prawitz, D. (2019a), "The fundamental problem of general proof-theory", in *Studia Logica* 107 (1), pp. 11-29.
Prawitz, D. (2019b), "The seeming interdependence between the concepts of valid inference and proof", in *Topoi*, 38, pp. 493-503.
Prawitz, D. (2024), "The interdependence between the concepts of valid inference and proof revisited", in A. Piccolomini d'Aragona (ed. by), *Perspectives on deduction: contemporary studies in the philosophy, history and formal theories of deduction*, Springer, pp. 21-37.
Ranta, A. (1994), *Type theoretical grammar*, Oxford University Press.
Rathjen, M. (2005), "The constructive Hilbertian program and the limits of Martin-Löf type theory", in *Synthese*, 147, 1, pp. 81-120.
Schiemer, G. & Giovannini, E. (2024), *Hilbert metatheory revisited: categoricity and interpretability*, Forthcoming.
Schiemer, G. & Reck, E. H., "Logic in the 1930s: type theory and model theory", in *Bulletin of symbolic logic*, 19(4), pp. 433-472.
Schroeder-Heister, P. (1991), "Uniform proof-theoretic semantics for logical constants (abstract)", in *Journal of symbolic logic*, 56, p. 1142.
Schroeder-Heister, P. (2006), "Validity concepts in proof-theoretic semantics", in *Synthese*, 148, 525-571.
Sieg, W. (2013), *Hilbert's programmes and beyond*, Oxford University Press
Smorynski, C. (2008), "Hilbert's programme", in E. Menzler-Trott, *Logic's lost genius: the life of Gerhard Gentzen*, AMS-London Mathematical Society, pp. 291-342.

Sundholm, G. (1994), "Vestiges of realism", in B. McGuinness & G. Oliveri (eds), *The philosophy of Michael Dummett*, Springer, pp. 137-165.

Sundholm, G. (2001), "A plea for logical atavism", in O. Majer (ed. by), *The Logica yearbook 2000*, Philosophia, pp. 151-162.

Sundholm, G. (2009), "A century of judgement and inference: 1837-1936. Some strands in the development of logic", in L. Haaparanta (ed. by), *The development of modern logic*, Oxford Academic, pp. 263-317.

Sundholm, G. (2019), "The neglect of epistemic consideration in logic: the case of epistemic assumptions", in *Topoi*, 38, pp. 551-559.

Tarski, A. (1956a), "On the concept of logical consequence", in A. Tarski, *Logic, semantics, metamathematics. Papers from 1923 to 1938*, Oxford University Press, pp. 409-420.

Tarski, A. (1956b), "The concept of truth in formalized languages", in A. Tarski, *Logic, semantics, metamathematics. Papers from 1923 to 1938*, Oxford University Press, pp. 158-278.

The Univalent Foundations Program (2013), *Homotopy Type Theory: univalent foundations of mathematics*, https://homotopytypetheory.org/book

Troelstra, A.S. & van Dalen, D. (1988), *Constructivism in mathematics*, voll. I II, Elsevier.

von Plato, J. (2008), *From Hilbert's programme to Gentzen's programme*, in E. Menzler-Trott, *Logic's lost genius: the life of Gerhard Gentzen*, AMS-London Mathematical Society, pp. 367-403.

Wang, H. (1996), *A logical journey: from Gödel to philosophy*, MIT Press.

Multi-field as a determinable

Davide Romano
University of Verona
davide.romano@univr.it

1. Introduction: the status of the wave function in quantum mechanics

Quantum mechanics (QM) is a rather peculiar theory: on the one hand, it is a very successful theory and no one doubts that it grasps something *true* about the nature of the microscopic world, on the other hand, however, the theory is metaphysically obscure, as the link between the formalism and the ontology of systems is not clear. As a result, different interpretations have proposed different ways to connect the formalism with the ontology. The standard interpretation of quantum mechanics (SQM, the theory presented in QM textbooks)[1] is manifestly an operational theory, and in doing so it renounces to provide a realist description of systems. Literally taken, SQM indicates the spectrum of possible measurement results (eigenvalues) and their relative probability distribution. This theory works very well in practice, but it does not provide an ontology of systems independently from measurement. For example, a plane wave cannot be interpreted as physical wave, since the ontology of SQM concerns the eigenvalues and not the wave function per se and, most importantly, the wave function cannot be generally defined as a classical field in 3D space. This is consistent with the standard interpretation, where the wave function is a probability amplitude (not an ontological entity) and the ontology of the theory concerns the eigenvalues/measurement outcomes, but it leaves the question about the nature of quantum systems basically unanswered.[2]

[1] E.g. Landau-Lifshitz (2013); Sakurai & Napolitano (2020); Shankar (2012).
[2] For an analysis of the ontology of standard quantum mechanics, see e.g. Ballentine (2014), Bowman (2008), Maudlin (2019), Norsen (2017).

Davide Romano, “Multi-field as a determinable”, in Claudio Ternullo and Matteo Antonelli, *Artificial minds, realism and evidence in science. Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)*, pp. 245-275

Università degli Studi di Urbino Carlo Bo
http://isonomia.uniurb.it/epistemologica

In order to overcome these limitations imposed by SQM, a certain number of non-standard interpretations—such as the Many Worlds Interpretation (MWI)[3] and Relational Quantum Mechanics (RQM)[4]—as well as non-standard theories—such as the Ghirardi-Rimini-Weber (GRW)[5] and the de Broglie-Bohm (dBB)[6] theories—have been proposed in the last decades.[7] All these approaches attempt to retrieve a realist description of quantum systems[8] while leaving the empirical predictions of SQM intact.[9] It is a hard attempt, and in fact one that originated much controversy in the philosophical literature: controversy concerning the best theory to adopt, controversy on the correct metaphysical interpretation for each of the non-standard theories. Leaving aside the former issue, the latter one is closely connected to the interpretation of the wave function.

While SQM and RQM are clear on the status of the wave function, as in both theories the wave function is purely instrumental,[10] this question is genuinely open in the MWI, GRW and dBB theories. Since all of these theories aim to provide a realist account of quantum mechanics, the wave function also seems to take more than just an instrumental role. In particular,

---

[3] Wallace (2012).

[4] Rovelli (1996).

[5] Ghirardi *et al.* (1986).

[6] Bohm (1952); Bohm & Hiley (1993).

[7] I distinguish between non-standard interpretations (MWI, RQM), which do not change the formalism of quantum mechanics (even though the MWI does not include the collapse postulate and may well be considered a non-standard theory as well) and non-standard theories, which do change the formalism of the theory by modifying the Schrödinger's equation (GRW theory) or the definition of the state (dBB theory).

[8] The situation is actually more nuanced: the GRW and Everett theories have been originally proposed as solutions of the measurement problem and unification of the micro and macro regime, whereas the dBB theory has been proposed not to solve the measurement problem but to provide a realist account of quantum systems. Moreover, this theory was originally proposed by Louis de Broglie in 1927 during the Solvay conference, so it is historically as old as the Copenhagen interpretation (see e.g. Baccciagaluppi & Valentini (2009) for the early history of the de Broglie's theory).

[9] All non-standard theories must recover the empirical results of quantum experiments, which are successfully described by standard quantum mechanics.

[10] I list Relational Quantum Mechanics among the *realist* interpretations because, even if this theory is instrumental on the wave function, it describes objective *relative facts,* which are the result of any interaction between systems and do not depend on measurement operations. However, the debate on the metaphysics of RQM is still open in the literature and one may provide arguments to interpret RQM as an instrumental theory *tout court*. While I would rather defend the former option on this point, the latter one is not excluded.

three different interpretations of the wave function in these theories have been proposed so far:

1. The nomological interpretation[11]
2. The 3N-D field interpretation[12]
3. The multi-field interpretation[13]

Even if it is common to discuss these approaches on similar grounds, it must be noticed that they have different areas of applicability, i.e. not all these interpretations can be consistently applied to all theories. In particular, the *nomological view* and the *multi-field approach* have a more restrictive application than the *3N-D field interpretation*. While the latter can be applied to MWI, GRW and dBB theories, the nomological interpretation can be applied only to theories with a primitive ontology, that is, to dBB, GRW_m and GRW_f[14] theories. The multi-field approach is even more restrictive, as it can be consistently applied only to the dBB theory (either in the first-order Bohmian mechanics or in the second-order Bohm's 1952 theory).[15] It may be useful to summarize these remarks in the following table (1.1):

Interpretation of the wave function	**Areas of applicability**
3N-D field	MWI, BM (1st-order), Bohm's theory (2nd-order), GRW_0, GRW_m, GRW_f
Nomological view	BM (1st-order), GRW_m, GRW_f
Multi-field	BM (1st-order), Bohm's theory (2nd–order)

Table 1.1

[11] Goldstein & Zanghì (2013).

[12] Albert (2013); Ney (2021). I use the term "3N-D field interpretation" rather than "wave function realism" as there are different ways in which one can be realist on the wave function. For example, the multi-field approach is definitely a realist interpretation of the wave function, but it differs substantially from Albert's and Ney's wave function realism.

[13] Forrest (1988); Belot (2012); Hubert & Romano (2018); Romano (2021a).

[14] GRW_m and GRW_f stands, respectively, for "GRW with mass-density" and "GRW with flashes".

[15] The reasons for this restriction will be clear in sect. 3.3. At the current stage there are informal attempts to extend the multi-field approach to other contexts, such as the GRW theory. While I remain skeptical that such extension can be consistently done, it will certainly be a positive result if these attempts will turn out to be eventually possible.

This paper does not want to enter in the debate concerning the best interpretation of the wave function, as (I am fairly convinced that) this is eventually left to personal preferences and perspectives. The aim of the present paper is instead more humble and, at the same time, more concrete: I want to provide a precise *metaphysical characterization of the multi-field* in terms of the *determinable-based account of metaphysical indeterminacy*.

The paper has the following structure: in (sect. 2) I review the interpretation of the wave function in standard QM; in (sect. 3) I describe the nomological view (sect 3.1), the 3N-D field interpretation (sect. 3.2) and the multi-field approach (sect 3.3). In (sect. 4) I present the determinable-based account and, following Wilson (2013, 2017), connect it to metaphysical indeterminacy. Finally, in (sect. 5) I describe the multi-field as a novel physical entity in terms of the determinable-determinate account. In (sect. 6) I draw some connections between the multi-field as determinable and relevant features of Bohm's theory. Conclusions are given in (sect. 7).

2. The wave function in standard quantum mechanics

In SQM the wave function of a system represents completely the state of the system but it has no ontological significance: it is rather an instrument for computing (given a certain observable) the spectrum of possible eigenvalues and their probability distribution. The meaning of the *state* is one of the most relevant differences between quantum and classical mechanics. In classical mechanics the system's state (initial position and velocity) has a direct reference to the system's ontology: the state represents the position and velocity of the system at the initial time. The classical state has therefore a double role, representational and ontological: it specifies the degrees of freedom needed to compute the evolution of the system (representational role) and, at the same time, it refers to a concrete system in space and time (ontological role). In SQM, instead, the state is just representational: the wave function represents the complete information on the system's state, i.e. the information needed to compute the evolution of the system (via the Schrödinger's equation), but it does not have a direct link with the system's ontology: we do not know what kind of system the wave function represents, e.g. if the system *is* a particle, or a wave, or a "wave-particle" or just a novel entity.

The physical meaning of the wave function in SQM is given by its absolute square $|\psi|^2$, which is interpreted as a probability density (Born's statistical interpretation). Consequently, the integral of this quantity: $\int |\psi|^2 dx$

gives the probability to obtain specific eigenvalues for specific observables. So construed, SQM is a theory about the possible measurement results (eigenvalues) of different observables. The eigenvalues are obtained from the collapse of the wave function, which is a postulate of the theory: in a measurement of the observable A, represented by the Hermitian operator $\hat{A}$, the wave function collapses instantaneously in one of the eigenstates $|a_k\rangle$ of $\hat{A}$ and the measurement result is mathematically represented by the eigenvalue a_k associated to $|a_k\rangle$.[16]

We note that the eigenvalue cannot be assigned to the system before the collapse takes place, that is, before and independently of the measurement process. This means that a quantum superposition (e.g. a superposition of different locations in the two-slit experiment or a superposition of "spin-up" and "spin-down" in the singlet state of the electron) cannot be interpreted as a superposition of different eigenvalues as the collapse has not yet occurred.[17] In such contexts, we must refrain to associate to the system a superposition of real-existing classically incompatible values since, according to the postulates of quantum mechanics, we can associate physical values to a quantum system only through eigenvalues and we cannot associate any eigenvalue before a measurement has taken place. We should distinguish instead between *representational* and *ontological* capacity of the wave function, where the former is the ability to mathematically represent the system and the latter the ability to indicate which kind of entity the system described by the wave function is. While SQM succeeds in the former task, it leaves the question about the ontology (latter task) basically unanswered. This aspect is unsatisfactory if we want to provide an ontology for quantum systems and is the main reason to look at the non-standard theories introduced above. Therefore, we now turn to the metaphysical analysis of the wave function in such theories.

---

[16] The Hermitian operator is defined by the action on its eigenstates: $\hat{A}|a_k\rangle = a_k|a_k\rangle$.

[17] The only exception being if the state is an eigenstate of a given observable, according to the eigenvalue-eigenstate link. In this case, the state does not change during the measurement process, so the collapse does not apply and we can assign a specific eigenvalue to the state independently from the measurement.

3. The wave function in non-standard quantum mechanics

Differently from SQM, the wave function can take an ontological meaning in MWI, GRW and dBB theories.[18] In these theories the spectrum of the metaphysical interpretations of the wave function is much greater: it can have an *instrumental role* (as commonly assumed in the GRW physics community), or a *nomological role* (as in the nomological view, where it guides the motion of the primitive ontology) or an *ontological role*, where it can represent a physical entity (as in the 3N-D field interpretation and the multi-field approach) or just patterns in three-dimensional space (as in Wallace's MWI).[19] The debate is open and there is no consensus on what the wave function *is* or *represents* in these theories. In the next subsections I will present the three major interpretations that have been proposed so far in these contexts, namely the *nomological view*, the *3N-D field interpretation* and the *multi-field* approach.

3.1. The nomological interpretation

The nomological interpretation has been originally proposed by Goldstein & Zanghì (2013), even though some traces of this interpretation date back to Hiley and Bohm's (1993) notion of *active information*. According to this interpretation, the wave function in the dBB theory is a nomological entity, i.e. a mathematical object that has no ontological counterpart but that is necessary to describe the evolution of the system. The analogy is with the Hamiltonian function in classical mechanics: as the Hamiltonian function (mathematically represented in phase space) "guides" the motion of the particles in 3D space, the wave function (mathematically represented in configuration space) "guides" the motion of the Bohmian particles in 3D space via the guiding equation:[20]

$$\dot{q} = \frac{\hbar}{m} Im\left(\frac{\nabla\psi}{\psi}\right) \qquad (3.1)$$

[18] I refer specifically to MWI since the original Everett's theory ("relative state formulation" of QM) is much closer in spirit to relational quantum mechanics: it does not postulate the existence of branching parallel worlds, it is observer-dependent and it describes sequences of "records" relative to the observers in line with RQM's relative facts.
[19] Wallace (2010).
[20] For simplicity, I write the guiding equation for spinless particles.

That is: given $\psi(x, t)$ we can compute the velocity and so the trajectories of the Bohmian particles in the same manner as we can compute the trajectories of classical particles in Hamiltonian mechanics from $H(x, v)$. This interpretation is quite attractive as it dissolves the problems linked to the multi-dimensionality of the wave function (the fact that, for an N-particle system, the wave function is defined in 3N-D space rather than in 3D space), but it also faces important issues. For example, a typical nomological entity (like the classical Hamiltonian) is not time-dependent and is not contingent (i.e. it does not depend on the boundary conditions), while the wave function is contingent and (generally) time-dependent. In order to solve this problem, Goldstein and Zanghì posit that only the wave function of the universe--the *Universal Wave Function* (UWF)—has an ontological significance. The UWF, as a solution of the Wheeler-de Witt equation, is supposed to be unique and time-independent.

However, this creates a further problem: in quantum mechanics we typically assign wave functions to (isolated) subsystems, never to the universe as a whole, exception made for quantum cosmology. That the UWF is the only wave function that counts from the ontological point of view is a metaphysical postulate. In addition, the very definition of UWF is not obvious: the wave function of the universe, if it exists, may well be represented by a factorized state between different (effective) wave functions, as it is plausible to assume that not all the regions and parts of the universe have previously interacted with each other, forming a unique entangled state. Furthermore, even leaving aside the problems associated to the universal wave function, there is a fundamental structural asymmetry between a nomological entity like the Hamiltonian, which has a *bottom-up* structure, i.e. it is built "from below" by the sum of the kinetic and potential energy of the particles, and a typical wave function, which has a *top-down* structure, i.e. it is derived as a solution of a dynamical equation (the Schrödinger's equation), as the electromagnetic field is a solution of the Maxwell's equations.[21]

3.1.1. Active information vs nomological view

The idea of *active information* proposed by Bohm & Hiley (1987: 327-328) shares some common features with the nomological interpretation: according to Bohm & Hiley, the wave function is a sort of information pool that guides the motion of the particles in the same manner as the electromagnetic waves produced by a remote guide the motion of a radio-controlled boat. As the

[21] See e.g. Romano (2021a, sect. 2).

electromagnetic waves transmit to the boat the information about its future movement (so the boat moves according to the transmitted information), the wave function transmits to the Bohmian particles the information on their future trajectory (so the Bohmian particles move according to the information transmitted by the wave function). The idea of active information may be seen as a precursor of the nomological interpretation, as it is a first attempt to regard the wave function as a non-material object (information pool, nomological entity) which guides the motion of a material object (Bohmian particles). The way it transmits this information to the particles is different in the two cases, but the general schema (action of a non-material entity to a material entity) is the same. Not really "what the Doctor orders", but what the information transmits.

3.2. 3N-D field interpretation

The *3N-D field interpretation*, most commonly known as *wave function realism*, has been originally proposed by Albert in two papers (1996, 2013). The idea is to interpret the wave function as close as possible to the role it plays in the quantum formalism: the wave function looks like a physical field, for it is the solution of a dynamical equation and, like a field, it assigns specific values to each point of the space on which it is defined. Since the wave function is defined on the system's configuration space, it assigns values to each point of that space, not to points of three-dimensional space.

Based on these features of the quantum formalism, Albert proposes to regard the wave function as a *physical field in configuration space.* As a classical field (e.g. the electromagnetic field) assigns real values to points of 3D space, the wave function assigns complex values to points of configuration space. According to this view, the wave function is analogous to a classical field, the only difference being that it is defined in configuration space rather than in three-dimensional space and that assigns complex values rather than real values to each point of its domain. There is however an important consequence: since the wave function is a physical field in configuration space, the latter must be recognized as the fundamental physical space of quantum mechanics. And since quantum mechanics is more fundamental than classical mechanics,[22] configuration space must be seen as

[22] A fair consensus has been reached in the literature that environmental decoherence plays an important role in the quantum-to-classical transition (Joos *et al.* (2013), Schlosshauer (2007, 2019); Zurek (2002)). According to this picture, the classical world emerges from the quantum world when quantum systems interact strongly and continuously with the external

the fundamental space of physics *tout court*. This is why this position is better defined as *configuration space realism* rather than wave function realism: one may be realist on the wave function without endorsing the fundamentality of configuration space (as it happens e.g. in the multi-field approach).[23]

If configuration space is fundamental, then the configuration of the Bohmian particles must be also placed on that space rather than on three-dimensional space. Following this reasoning, Albert reduces the (somewhat illusory or emergent) configuration of particles in 3D space to a "marvelous point" in 3N-D space. The marvelous point solves the communication problem[24] between the wave function and the Bohmian particles as they are both placed in the same space, but it does not help with the "perception problem", i.e. the problem to understand why we perceive the macroscopic world as three-dimensional even though the fundamental space is the configuration space. Solutions to the perception problem have been proposed by Albert (2013) and more recently by Ney (2021). Albert argues that it is the structure of the Hamiltonian that decomposes configuration space into sets of three-dimensional coordinates, giving the impression that this is the space where particles move and interact with each other, and eventually giving the impression that we live in 3D space. Ney's argument relies instead on the role of symmetries in quantum mechanics. In particular, she notes that, even though the fundamental space is configuration space, important symmetries of quantum mechanics are retrieved only when we represent systems in 3D space. In both cases, however, an open question remains on how these mathematical structures (Hamiltonian, symmetries) can affect our perception to live in a 3D world.

---

environment. The classical world in this picture is emergent or derivative with respect to the quantum world. See, however, Romano (2022) for a critical assessment of environmental decoherence in the standard context. A dissident voice on the importance of environmental decoherence for the classical limit is Ballentine (2008).

[23] Furthermore, we note that configuration space, differently from three-dimensional space, does not have a fixed number of dimensions, as the latter depends on the number of degrees of freedom of the system under analysis. In particular, given an N-particle entangled state, the system's configuration space has a number of dimensions $3N$, where N is the number of particles composing the system. The number of dimensions therefore vary from system to system. This reflects the derivative character of configuration space with respect to three-dimensional space and, to my opinion, it is not a firm ground to assess the fundamentality of configuration space over three-dimensional space.

[24] The problem of communication arises when the wave function and the Bohmian particles "live" on different spaces (e.g. Callender 2015) and can be summarized as follows: how does the wave function (object in 3N-D space) guide the Bohmian particles in 3D space?

3.3. The multi-field approach: the wave function as a new entity in 3D space

In the multi-field view, the wave function is the mathematical representation of a *multi-field*, which has to be regarded as a novel physical entity in 3D space. The idea of the multi-field comes originally from the notion of "polywave" proposed by Forrest (1988). Forrest interprets the wave function in SQM as a "polywave", that is, as a multiple assignment of field values for any ordered set of position coordinates. The notion of polywave has been then revisited and inserted in the context of Bohm's theory by Belot (2012), who names it "multi-field". However, Belot dismisses quickly the multi-field idea, principally because of the non-validity of the action-reaction principle (while the multi-field acts on the Bohmian particles, the latter do not act back on the former).[25] After this first attempt, the multi-field approach has been further developed and defended in Hubert & Romano (2018) and more recently in Romano (2021a).

As we saw before, the wave function looks like a field as it defines a specific value for each point of the space on which it is mathematically defined, yet these values are associated to points of configuration space and (for N-particle entangled states) they cannot be reduced to an assignment of pre-existing values associated to points of 3D space. In other words, the wave function generally assigns a continuous distribution of complex values to each point of the system's configuration space. The idea of the multi-field is to interpret such distribution of values in configuration space as the mathematical representation of a novel physical entity in 3D space. More precisely, the assignment of definite values in configuration space is interpreted not as a classical field in configuration space, but rather as a novel kind of object in 3D space. The new object is the multi-field.

Even though the wave function cannot assign pre-existing, determinate values to each point of 3D space, a projection from configuration to three-dimensional space can always be done, as configuration space is literally the space of possible configurations of particles in three-dimensional space. We can illustrate this idea with the following example. Consider a system of two point-particles in 3D space represented by the coordinates $p_1(x_1, y_1, z_1)$ and $p_2(x_2, y_2, z_2)$: we can represent the 2-particle system as two discrete particles in 3D space or, equivalently, as a single particle in 3N-D space:

[25] A response to Belot on the back-reaction problem is given in Romano (2021a, sect 6.3).

Two-particle system in configuration space

$$p_{1,2}(x_1, y_1, z_1, x_2, y_2, z_2) \qquad (3.3.1)$$

We note, however, that the position coordinates of the two particles in the single particle representation are ordered: the first set of three coordinates (x_1, y_1, z_1) represent particle p_1, the second set (x_2, y_2, z_2) represent particle p_2. Configuration space is built from the configuration of particles in 3D space: we define p_{12}, a single point in 3N-D space, from the configuration of two particles p_1 and p_2 in 3D space. Since configuration space is derivative from three-dimensional space, the latter can be safely viewed as the fundamental physical space, i.e. the arena where systems exist and interact with each other. However, we are still left with the initial problem: the wave function assigns precise values to points of configuration space, not to points of 3D space. Let us clarify this point in the context of Bohm's theory.

Consider a 2-particle entangled state. In Bohm's theory this system is represented by a six-dimensional wave function $\psi_{1,2}(x_1, x_2, \ldots, x_6)$ and by the actual particles' configuration q_{tot} composed of two point-particles, q_1 and q_2, having exact locations and mathematically represented by the position coordinates in 3D space: $q_1(x_1, y_1, z_1)$, $q_2(x_2, y_2, z_2)$. The Bohmian system is thus represented by the state: $(\psi(x), q_{tot})$.

The wave function specifies a map from configuration space to the complex numbers:

$$\psi_{1,2}(x_1, x_2, x_3, x_4, x_5, x_6) \to c \qquad (3.3.2)$$

If we leave the interpretation at this stage, we have the original idea of Forrest's polywave (the wave function assigns a complex value to any ordered N-tuples of points), but in Bohm's theory the wave function is always accompanied by the actual configuration q. When we insert the particles' configuration into the wave function:

$$\psi(q_1, q_2) = \psi(x_1, y_1, z_1, x_2, y_2, z_2) \qquad (3.3.3)$$

the wave function assigns a complex value to the two discrete points, (x_1, y_1, z_1) and (x_2, y_2, z_2), in 3D space corresponding to the exact location of the Bohmian particles q_1 and q_2. In general, for an N-particle system, the wave function assigns a complex value to the N-tuple of points corresponding to the actual particles' configuration. The result is not a classical field, as it assigns simultaneously a specific complex value (c) to an N-tuple of points in 3D space (in our case at the two points (q_1, q_2)) and the value is not pre-assigned but depends on the position coordinates of the particles composing the configuration q. Under this view, the wave function is thus a new kind of physical field--a multi-field--which assigns specific field values to N-tuples of points corresponding to the exact location of the Bohmian particles.

The multi-field so described can be thought as a generalization of a classical field: while a classical field (e.g. the electromagnetic field) assigns a determinate value to any point of 3D space, the multi-field assigns a determinate value only to N-tuple of points, corresponding in Bohm's theory to the actual position of the particles. For example, given a wave function of the type:

$$\psi(x, y) = A\cos(xy) \qquad (3.3.4)$$

and configuration $q = (q_1, q_2)$, with A a normalization constant, the multi-field assigns a determinate value:

$$\psi(q_1, q_2) = A\cos(q_1 q_2) \qquad (3.3.5)$$

in correspondence of the two points $q_1$and q_2 occupied by the Bohmian particles. The determinate is computed by evaluating the 2-particle wave function at the points $x = q_1$ and $y = q_2$.

Note that the multi-field cannot be thought of as a continuous distribution of (determinate) values, differently from a classical field. While a classical field defines a determinate value at any point, the multi-field defines a determinate value only at those points where the Bohmian particles are located, leaving all the other (empty) points with indeterminate values. This constitutes a discontinuity in the field, and a primary difference with a classical field.[26]

The multi-field assigns a determinate complex value to a given N-tuple of points at any instant, corresponding to the exact location of the Bohmian particles at that instant. If complex values may sound unphysical, we note that the wave function can be reduced to two (coupled) real-valued functions, corresponding to the amplitude $R(x, t)$ and phase $S(x, t)$ of the wave function written in polar form: $\psi(x, t) = R(x, t)e^{\frac{i}{\hbar}S(x,t)}$. Consequently, the complex-

[26] However, I do not think this is a kind of discontinuity that should be handled mathematically, since the multi-field represents a novel entity with respect to a classical field, one that is fundamentally characterized by discontinuity. This fundamental discontinuity is reflected in the metaphysical characterization of the multi-field as determinable (sect. 5).

valued multi-field can be reduced to two (coupled) real-valued multi-fields associated to $R(x,t)$ and $S(x,t)$.[27]

In practice we do not know the exact location of the Bohmian particles but we know that, given a system with wave function ψ, the actual configuration is statistically distributed according to the Born's rule: $\rho(q) = |\psi(x)|^2$. This postulate is known as *quantum equilibrium* and guarantees that the de Broglie-Bohm theory is empirically equivalent to standard QM. Since a Bohmian system is defined, at any time, by a unique actual configuration q_{t^*}, the multi-field assigns, at any time, a unique and specific value to the N-tuple of points $(x_1, \ldots, x_N)$ corresponding to $q_{t^*} = (q_1, q_2, \ldots, q_N, t^*)$:

$$\psi(q_1, q_2, \ldots, q_N, t^*) \rightarrow c \qquad (3.3.6)$$

At any time, the multi-field assigns a complex value c to the N-tuples of points in three-dimensional space: $q_1, q_2, \ldots, q_N$, corresponding to the exact location of the Bohmian particles. Differently from Forrest's polywave, in Bohm's theory the multi-field assigns a unique determinate value associated to the configuration q at any instant of time. Even if we do not know the precise location of the particles (but only that they are distributed according to the Born rule), this is as a matter of fact an epistemic ignorance and does not affect the ontology described so far: even if the actual configuration is epistemically unknown, still *the Bohmian particles have an exact location in 3D space,* so the ontology of the multi-field is unambiguously determinate.

The multi-field can be regarded as a generalization of a classical field.[28] Whereas a classical field assigns a specific value to each point of 3D space, the multi-field assigns a non-local value to N-tuples of points of 3D space. I say "non-local" as the specific value assigned at one point (corresponding to the exact location of one particle of the configuration) depends non-locally (i.e. simultaneously at a distance) on the exact location of all the other particles of the configuration. It remains a problem, however: the multi-field does not specify any determinate value to the *empty points*, i.e. all points in the domain of the wave function that are not occupied by the Bohmian

[27] This approach has been proposed in Romano (2021a). Regarding the amplitude and phase of the wave function as multi-fields provides physical support to the quantum potential Q and quantum force F_Q, which enters in the definition of the quantum Newton's law: $F_C + F_Q = m\ddot{q}$. In fact, the quantum force is generated by the quantum potential: $F_Q = -\nabla Q$ and the latter is generated by the amplitude of the wave function: $Q = -\frac{\hbar^2}{2m}\frac{\nabla^2 R}{R}$.

[28] The multi-field as determinable is a novel physical entity with respect to a classical field in the same manner as the classical field is a novel physical entity with respect to the Newtonian force.

particles. This raises a prompt objection: how do we know that the multi-field includes also the empty points (i.e. that is an entity distributed in space rather than a relation between N points) if the latter are not associated to any determinate value? In order to solve this problem, I propose the following hypothesis: the multi-field is a *determinable* representing metaphysical, object-level, indeterminacy (QI). The multi-field represents an indeterminate state of affairs (Wilson 2013), yet it is ontologically as real as a classical field or a classical point-particle. Before expanding on this point, we introduce in the next section the determinable-determinate account and its link with metaphysical indeterminacy.

4. Metaphysical indeterminacy

Metaphysical indeterminacy (MI) is the idea that there is a state of affair of the world that is indeterminate and that such indeterminacy is intrinsic of the world itself. Such indeterminacy is therefore different from epistemic indeterminacy (coming from the lack of knowledge) or semantic indeterminacy (coming from vagueness or ambiguity of language). We may say: *semantic indeterminacy* comes from an imperfect correlation between the language and a determinate world, *epistemic indeterminacy* comes from an incomplete knowledge of the determinate world, *metaphysical indeterminacy* is the acknowledgment that the world itself is indeterminate.

Metaphysical indeterminacy divides into two main accounts: *meta-level* and *object-level* MI (Wilson 2013). The former is represented by *metaphysical supervaluationism (*Barnes 2010; Barnes & Williamson 2011); the latter by the *determinable-based* or *determinable-determinate* account (Wilson 2013, 2017). Following Wilson (2013) and Calosi & Mariani (2021), the difference between these two accounts of metaphysical indeterminacy is that:

> [A]ccording to the former [metaphysical supervaluationism] it is indeterminate which determinate state of affairs obtains (SOA), whereas according to the latter [determinable-based account] it is determinate that an indeterminate SOA obtains. [Calosi & Mariani (2021: 8)]

4.1. Supervaluationism

Metaphysical supervaluationism can be roughly summarized by the following quote by Barnes (2010: 622):

> It's perfectly determinate that everything is precise, but [...] it's indeterminate which precise way things are.

Calosi & Mariani (2021: 9) describe how supervaluationism can be applied to quantum mechanics, in particular how a superposition state can be described using supervaluationism:

> In general, consider a system S in state $|\omega\rangle = c_1|\psi\rangle + c_2|\phi\rangle$. There is MI because there are two admissible precisifications, the SOA that ψ and that ϕ respectively, and it is indeterminate which one is the case. That is, superposition indeterminacy boils down to indeterminacy about which term of the superposition obtains.

However, we can safely dismiss supervaluationism from our analysis for two reasons. First, supervaluationism does not seem to capture the characteristics of quantum mechanics. A superposition state is not a state in which the two eigenstates (precisifications, in this case) are determinate but we do not know which one obtains. This description does not capture the essence of a quantum superposition, in which all eigenstates (in the case above: $|\psi\rangle$ and $|\phi\rangle$) concur to the description of the behavior of the system, represented by the state vector $|\omega\rangle$ with different probability associated to each state (given by the absolute square of the associated coefficient). A state of affairs in which all the eigenstates of a superposition are equally determinate, as proposed by supervaluationism, would fail to generate the typical quantum interference that we observe in quantum experiments. For example, in the double-slit experiment with electrons or photons, the interference pattern that is progressively generated on the screen can be accounted for only considering constructive and destructive interferences between the two components between the slits and the screen, and the latter can be accounted for only considering different amplitudes between the interfering components. This tension is reported in Calosi & Mariani (2021: footnote 17): "we should note that the straightforward application raises questions on how to understand the coefficient c_1 and c_2 in the quantum state".

Furthermore, supervaluationism so defined seems to collapse into epistemic indeterminacy. If the world is totally precise and composed of multiple determinates, then it is just a matter of convention or lack of knowledge which one of these determinates represents the actual world. For example, this is how Darby (2010: 235) applies metaphysical supervaluationism to the Schrödinger's cat paradox:[29]

> [There is] a suggestive parallel between the terms in the superposition and the idea [...] of precisifications. One of the terms in the superposition [...] is a term where the cat is alive, the other is not; that is reminiscent of multiple ways of drawing the extension of 'alive', on some of which 'the cat is alive' comes out true, on some, false.

We see that this description does not seem to capture the essential features of the paradox: the cat in the box (before a measurement is performed) is in a quantum superposition of being alive and dead, as the cat is in an entangled state with the radioactive material in the box, which is represented by a coherent superposition of two definite states, being decayed and not decayed (more precisely, the radioactive material is represented, in general, by a decreasing exponential function that describes the probability amplitude of the radioactive decay as a function of time). According to SQM the result of a measurement on the state of the cat will describe a determinate state of affair, but such determinate SOA cannot be ascribed to the eigenstates associated to the cat in the box before a measurement is performed. If supervaluationism does that, then it would be in conflict with standard quantum mechanics. A more promising approach is the determinable-based account introduced in the next section.

4.2. Determinable-based account

The *determinable-based account* of MI or *determinable-determinate* account has been introduced by Wilson (2013, 2017) and later applied to quantum indeterminacy (e.g. Wolff (2015); Calosi & Wilson (2018); Calosi & Mariani (2021); Fletcher & Taylor (2024)). The basic idea is that a state of affair is described by a property or an object represented by a *determinable* and a *determinate*, the two standing in a specific property-type relation. The determinable is more general and accounts for a spectrum of possible determinates, the determinate is a specific instance or realization or actualization of the determinable. This is, for example, how Wilson (2017) presents the determinable-determinate account:

> Determinables and determinates are in the first instance type-level properties that stand in a distinctive specification relation: the "determinable–determinate"

[29] The quote is reported in Calosi & Mariani (2021: 8-9).

> relation (for short, "determination"). For example, color is a determinable having red, blue, and other specific shades of color as determinates; shape is a determinable having rectangular, oval, and other specific (including many irregular) shapes as determinates; mass is a determinable having specific mass values as determinates.

We can report the cited examples of determinable-determinate relations in the following table (4.1):

Determinables	**Determinates**
Color	Red, blue, green, …
Shape	Rectangular, oval, …
Mass	Mass values $m_1, m_2,\ldots$

Table 4.1

As reported in the quote above by Wilson, a standard example of determinable-determinate relation concerns the property of *color.* By saying that an object is "colored" we specify a *determinable*: a property (the property of being colored) to which may correspond many specific instances (the spectrum of determinate colors). If we say that a certain (colored) object is "red" we specify a determinate (a specific, determinate color) for the given determinable (being colored). The determinable account is pyramidal: "red" is a determinate with respect to the determinable "being colored" but is a determinable with respect to different shades of red, such as "scarlet" or "vermillion".

Note that in all these examples the determinable does not exist independently from the determinate: it does not exist in the world a colored object without a specific color, or a shaped object without a determinate (regular or irregular) shape. We anticipate that the multi-field is a determinable of a different type: it is a determinable object which exists independently of its determinate. This is valid for the multi-field account presented here as well as for any application of the determinable-based account to quantum indeterminacy (e.g. Calosi & Wilson (2018)).

In physics, the determinable-based account has been applied to classical properties such as *mass* (of a classical system) and to quantum properties such as the *position* (Bokulich (2014)) and *spin* (Wolff (2015)) of a quantum system. There is however an important difference between the classical and quantum case. In the classical case, the determinable property is always accompanied by a determinate. Consider, for example, the mass of a table. We may say that the mass as determinable is the general property of a classical object (a table in this case) of having a mass. However, it does not exist a classical object that has a mass without having a specific mass value.

That is, in classical physics, the determinable (e.g. the mass property) is always accompanied by a determinate (a specific mass value). Same for colors or shapes: it does not exist a colored object without a specific color, or a shaped object without a specific shape.

This is not the kind of relation between determinable and determinate that we find in quantum mechanics. A quantum system that is in a superposition of eigenstates with respect to a certain observable does not have a specific value for that observable (before a measurement is performed). The observable in quantum mechanics can thus be represented by a determinable without a determinate. Two standard examples concern the position and the spin of a quantum system. Consider a 1-particle system represented by a plane wave:

$$\psi(x) = Ae^{\frac{i}{\hbar}px} \qquad (4.2.1)$$

where A is a normalization constant and p the momentum eigenvalue. This state indicates an equal probability distribution to find the particle in any point of the space in a position measurement:

$$P_x = |\psi(x)|^2 = |A|^2 \qquad (4.2.2)$$

Until a measurement is performed, the particle does not have an exact position in space, that is, the observable "position" has a determinable without a determinate. The example of plane wave is summarized by Bokulich (2014: 467) as follows:

> In quantum theory it is more typically the case that the degree to which the particle's momentum is specified allows us to say, for example, that the particles is located somewhere in this room, although it is not possible to say that is located in any particular point in the room. In other words, while it makes sense to talk about the particle having the property of position (that is to say the particles are in the room), that property cannot be ascribed a definite (precise) value.

To be precise, in standard quantum mechanics we cannot say that the particle "is located somewhere" before the measurement is performed, as this would imply an epistemic interpretation of quantum probabilities, which is in conflict with the standard interpretation. It would be more correct to say that the particle is located *nowhere* before the measurement. Consequently, in the determinable-based account of MI, a quantum system (in SQM) never has a determinate position if not in the precise instant of a position measurement.[30]

The example of spin as determinable is analyzed by Wolff (2015). The spin case is different from the position case as the latter is a scalar quantity while the former is a vectorial quantity. For this reason, the spin operator is always defined along a given direction, so we have three different operators: $\hat{S}_x, \hat{S}_y, \hat{S}_z$, which represent the spin operator, respectively, along the $x -$, $y -$ and $z -$ axis. Consider a $\frac{1}{2}-$spin particle (e.g. an electron): this particle has two possible eigenvalues or the spin $\left(+\frac{1}{2}; -\frac{1}{2}\right)$, respectively associated to the eigenstates "spin-up" $|\uparrow\rangle$ and "spin-down" $|\downarrow\rangle$. As the three operators $\hat{S}_x, \hat{S}_y, \hat{S}_z$ are mutually incompatible (so it does not exist a state that is an eigenstate simultaneously of two of these operators), when the electron has a determinate spin along a given direction, the spin along a different direction is represented by a superposition of two eigenstates and thus is not determinate.

From this analysis Wolff suggests that we must associate a determinable to each individual operator $\hat{S}_x, \hat{S}_y, \hat{S}_z$ and not to the spin property *tout* court. Furthermore, Wolff notes that while the determinable-based account describes well the relation between the spin property and the spin value along a given direction, it does not explain why the operators $\hat{S}_x, \hat{S}_y, \hat{S}_z$ are mutually incompatible, i.e. it does not explain why certain sets of determinables cannot have joint determinates (the same conclusion applies to all sets of non-commuting observables, such as e.g. position and momentum).

Finally, Wolff analyzes three approaches to correlate the spin as a determinable with metaphysical indeterminacy. The first is the one proposed by Funkhouser (2006: 566), according to which: "an amendment for the quantum level might be that every object instantiating a determinable also instantiates certain determinates to certain probabilities."[31] This approach however does not work in the case of spin: the determinate ("spin up" or "spin down") is always a well-defined value, while probabilities are associated to uncertainty about the specific measurement result, as reported by Wolff (2015: 384):

---

[30] The situation is even more tricky: position eigenstates are represented in SQM by Dirac delta functions, which are not solutions of the Schrödinger's equation. In practice, a quantum system is considered fairly localized in position when it is represented by a Gaussian or a well-localized state.

[31] The quote is reported in Wolff (2015, p. 383).

> [w]hat exactly the probabilities denote is of course controversial, but minimally they simply state the likelihood of finding a particle with spin value "up" and "down" respectively in a given direction. By adding in the probabilities, we simply seem to acknowledge the indeterminacy of the spin state, we don't eliminate it.

The second and third approaches are instead those proposed by Wilson (2013): we can think of a determinable as corresponding to the instantiation of multiple determinates ("glutty" MI) or to the instantiation of none of the determinates ("gappy" MI). In the first case, we should think of the different directions of the spin as different but complementary perspectives. The classical example is the iridescent feather where multiple determinates colors are realized with respect to different perspectives. In the case of the electron spin:

> [T]his would mean that we treat the determinate outcomes of spin measurement in different directions as different perspectives. Depending on which measurement we carry out, i.e. how we orient our Stern-Gerlach device, we will get a determinate z-spin up, say, or a determinate y-spin down, but it would be misleading to suggest that the electron only has a determinate z-spin or only a determinate y-spin. It is just that from the perspective (read: measurement) we have chosen, this is the determinate which is realized in our perspective. [Wolff (2015: 384)]

This approach also encounters a number of convincing objections. First, it looks very closely to an epistemic reading of quantum uncertainty, furthermore there is a difference between multiple determinates of the same determinable (e.g. spin up and spin down along x −direction) and multiple determinates associated to different determinables (e.g. spin up and spin down along the y −direction for the state $|\uparrow\rangle_z$) that does not seem to be correctly described by this approach. Building on this analysis, Wolff concludes (convincingly, in my opinion) that the approach considering "gappy" MI is the best one of the three:

> Of the three answers to the question of indeterminacy, then, the third seems to be the most promising. It is also the most radical revision of the determinables/determinate distinction, since it requires the instantiation of determinables without determinates. If that is to be possible, determinables have to be accepted into the ontology on equal footing with determinates.[32] [Wolff 2015, p. 385]

[32] Wolff also reports some reservation on this kind of approach, as it requires a radical revision of the current ontology: "It is not obvious that this is a price worth paying, given

This is very close to the idea proposed here and developed in the next section to characterize the multi-field as a novel physical object. This idea imposes a radical revision of the current ontology, but one that (likely) offers more clarity in the interpretation of quantum indeterminacy and, in general, in the interpretation of the quantum ontology. We note that applying the determinable-based account to the multi-field is a step further with respect to applying it to spin or position in SQM, as the multi-field is not a property of the system but (part of) the system itself in the de Broglie-Bohm theory. Under this novel approach, the determinable does not describe the properties of a system but the system itself: the wave function is interpreted as a multi-field and the system, represented by the wave function (and by the particles' configuration) is itself interpreted as a determinable, that is, as a new kind of object. In the next section we will expand on this point and characterize more precisely the multi-field as a determinable.

5. The multi-field as a determinable

The hypothesis presented here is that the wave function is the mathematical representation of a new physical entity, a *multi-field*,[33] which can be metaphysically characterized as a determinable, i.e. an object defined by properties without a determinate value. The multi-field is actually more complex than the determinable usually presented in the literature, as it assigns a determinate (a specific and unique complex value) to the N-tuple of points corresponding to the actual configuration of the Bohmian particles ($x_i = q_i$) and a determinable without a determinate to all the other points $x_i \neq q_i$. Following the determinable-based account of MI, the multi-field so defined implies ontological indeterminacy, i.e. it describes an indeterminate state of affairs in the world:

> Here I present an account on which what it is for there to be MI is for it to be determinate (or just plain true) that an indeterminate (imprecise) SOA obtains. I

how little the application of the determinables model seems to contribute to our understanding of quantum indeterminacy." (Wolff 2015, p. 385).

[33] The name "multi-field" is correct insofar we intend it as a true generalization of a classical field. This generalization is fully captured by the determinable-determinate account and provides an example of quantum indeterminacy. We note that the quantum indeterminacy introduced by the multi-field characterizes the entity itself, not the properties of the system. A Bohmian system has a definite position (specified by the actual configuration), a precise velocity (specified by the guiding equation), a precise acceleration (specified by the quantum Newton's law), yet the multi-field values at the empty points have an indeterminate value.

> more specifically suggest that the obtaining of an indeterminate SOA is profitably understood in terms of an object's having, on the one hand, a determinable property, but not having, on the other hand, a unique property that is a determinate of that determinable." [Wilson (2013: 360-361)]

Within the region R where the multi-field is well-defined (the projection of the wave function in 3D space), the determinable property is represented by the (complex) values that the multi-field assigns to each point of three-dimensional space. It is a determinable as (i) the value of each of these points $(x_i \neq q_i)$ is not determinate but, at the same time, (ii) a determinate is selected for any of these points once a particle is located at that point, i.e. when the initially empty point is included in the points corresponding to the actual configuration $x_i = q_i$. In other words, any empty point is characterized by a *set of possible (potentially infinite) multi-field values.* A specific value from this set is selected, however, when a particle of the configuration q is located at that point: the (originally empty) point will be so associated with a determinate, unique multi-field value.

This criterion of selection of the determinate is for some aspects similar to the way we select a value for a classical field, but for other aspects very different. Consider an electric field $\vec{E}(x,t)$ defined in the region Γ. This field assigns a specific value to any point $x \in \Gamma$ for any instant of time. The way we generally define a field value is associated to the indirect effect of the field on a charged test particle. For example, if we locate a test particle on the point $x_k \in \Gamma$ at time $t = t^*$, the particle will accelerate under the Lorentz force: $\vec{F}(x_k) = q\vec{E}(x_k, t^*)$. From the acceleration of the test particle we derive indirectly the existence of the electric field $\vec{E}(x,t)$ in that region. In the case of the multi-field we do not have test particles but we can divide the scheme between the wave function $\psi(x,t)$ and the Bohmian particles $q = (q_1, \dots, q_N)$. For simplicity, consider a two-particle state with wave function $\psi(x_1, x_2, t)$ and actual particle configuration $q = (q_1, q_2)$, defined in a one-dimensional potential box with length L. The points where the multi-field as determinable is well-defined correspond to the points where the wave function in 3D space is well defined, i.e. to all points: $0 \leq x \leq L$. Differently from the electric field, the multi-field does not assign a specific value to each point of the region $0 \leq x \leq L$, excluding the points $(x_1 = q_1; x_2 = q_2)$. Suppose, however, that we want to know the value of the multi-field associated to the (originally empty) point $x = \frac{L}{2}$. In this case, analogously to the case of the test particle, we can derive the (determinate) value of the multi-field at $x = \frac{L}{2}$ at the time $t = t^*$ by assuming to locate (as, in practice, we cannot control the position of Bohmian particles) one of the two particles of the configuration $q = (q_1, q_2)$ exactly at the point $x = \frac{L}{2}$. Suppose that we choose particle 1, represented by q_1: we thus consider the system $\psi(x_1, x_2, t^*)$ with particle configuration $q = \left(\frac{L}{2}, q_2\right)$. In this case, the multi-field will assign the (complex) determinate value:

$$\psi\left(\frac{L}{2}, q_2, t^*\right) = c \qquad (5.1)$$

to the couple of points $\left(\frac{L}{2}, q_2\right)$, that is, to the two points corresponding to the exact location of the Bohmian particles. We note from this example that the value of the multi-field at $q_1 = \frac{L}{2}$ is determinate but *non-local*, as it depends on the specific location q_2 of the other particle of the actual configuration.

The analogy here is that, as the test particle proves (indirectly) the existence of the electric field by the effect of the field on the particle, in a similar manner the effect on the Bohmian particle (the velocity via guiding equation or the acceleration via quantum Newton's law) proves (indirectly) the existence of the multi-field. In particular, we can compute the determinate multi-field value at any point of the region where the multi-field is well-defined by locating (hypothetically) a Bohmian particle of the actual configuration at that point. This process transforms a determinable (a set of infinite possible values) into a determinate (a specific complex value). There are, of course, two important differences in the classical and quantum case. First, the Bohmian particle is not a test particle. While in the case of the electric field we assume to put an external particle (test particle) to evaluate the value of the field, in the case of the multi-field we assume to put a particle of the actual configuration that composes the Bohmian system. Second, as mentioned before, the value of the multi-field at the point $x = \frac{L}{2}$ depends non-locally on the value of q_2, i.e. the position of particle 2. At any instant q_2 will be represented by a specific real number, and overall the multi-field will assign a unique determinate to the couple of points $\left(\frac{L}{2}, q_2\right)$. Yet, if we change the location of the second particle q_2 the multi-field value at $q_1 = \frac{L}{2}$ will also changes, as the multi-field assigns one specific value for the entire configuration: $\psi\left(\frac{L}{2}, q_2\right) = c$. Differently from the classical case, the determinate value of the multi-field at one point depends on the exact location of distant particles of the actual configuration. We may say that, differently from the classical case, the multi-field assigns a *non-local* determinate value to the N-tuple of points corresponding to the actual configuration of particles: $(x_1 = q_1, \dots x_i = q_i, \dots x_N = q_N)$. In this way, Bohmian non-locality (and quantum non-locality more generally) is implemented in the very definition of the multi-field. The multi-field as determinable can be naturally regarded as a *non-local beable*.[34]

The multi-field so defined is (plainly) a determinable: it describes an indeterminate but objective, ontologically real, state of affairs. This is exactly the state of affair associated to a determinable, as reported by Wilson (2013: p. 366):

> **Determinable-based MI**: What it is for a state of affairs to be MI in a given respect R at a time t is for the state of affairs to constitutively involve an object (more generally, entity) O such that (i) O has a determinable property P at t, and (ii) for some level L of determination of P, O does not have a unique level-L determinate of P at t.

In the *multi-field-as-determinable* account, the MI state of affair involves the object or entity "multi-field" M such that (i) M has a determinable property P at t, i.e. the multi-field values that it assigns at any empty point (excluding the points $x_i = q_i$) within the region where the wave function in 3D is well-defined and (ii) for any point $x_i \neq q_i$, M does not have a unique determinate of P at t. There are two levels L of determination: L_1, L_2. The first corresponds to the empty points within the multi-field region: $L_1(x_i \neq q_i)$, the second to the points of the actual configuration $L_2(x_i = q_i)$ For the level of determination L_1 there is no unique determinate of P: any point is associated with a set of possible multi-field values. For the level of determination L_2 there is instead a unique determinate: a specific complex value assigned to the N-tuple of points corresponding to the actual configuration $(x_1 = q_1, \dots, x_N = q_N)$.

The metaphysical indeterminacy implied by the determinable-based account can be characterized even more precisely. In fact, there are two ways in which a determinable can fail to have a unique determinate: either it has none, or it has more than one. The former case is termed "gappy" MI, the latter "glutty" MI. A standard definition is given in Calosi (2021: 11305):

> According to the Determinable Based Account (DBA) of metaphysical indeterminacy (MI), there is MI when there is an indeterminate state of affairs, roughly a state of affairs in which a constituent object x has a determinable property but fails to have a unique determinate of that determinable. There are

[34] On this point see also Hubert & Romano (2018, sect. 5).

different ways in which x might have a determinable but no unique determinate: x has no determinate—gappy MI, or x has more than one determinate—glutty MI.

The multi-field as determinable is a case of *gappy metaphysical indeterminacy*, as the determinable P fails to assign a determinate value at any point $x_i \neq q_i$. In conclusion, the multi-field as determinable is defined as a distribution of determinable-property P, that is, a set of possible complex values for each point within the region of 3D where the wave function is well-defined. At any empty point ($x_i \neq q_i$) corresponds a determinable without a determinate, however the point takes a determinate as soon as it is occupied by a particle ($x_i = q_i$). The specific value at that point will depend not only on the wave function but also on the exact location of distant particles that compose the actual configuration, so defining a *non-local determinate*.

6. Some remarks on the ontology of the multi-field and Bohm's theory

In this final section, I present some remarks on the metaphysics of the multi-field as determinable in connection with relevant features of Bohm's theory, in particular with the nature of non-locality, the guiding equation and the quantum equilibrium. These remarks are not intended to be complete, but they want to offer a suggestion on the metaphysical import of the multi-field view within the ontology of Bohm's theory.[35]

6.1. Multi-field as determinable and non-locality

From the discussion above, we notice that the multi-field as determinable implements Bohmian (and in general quantum) non-locality quite naturally, as the determinate depends at the same time on the precise location of all the Bohmian particles. Changing the position of one particle of the configuration instantaneously changes the determinate value that the multi-field assigns at that configuration. As suggested above, we can say that the determinate is *non-local*, according to this description. Consequently, the multi-field as determinable view accounts for the non-local correlations between distant particles (for N-particle entangled states) since the determinate value of the multi-fields depends instantaneously on the exact position of all the Bohmian particles of the configuration, no matter how distant they are. The Bohmian

[35] Thanks to an anonymous reviewer for inviting me to clarify these points concerning Bohmian non-locality, the guiding equation and the Born's probabilistic distribution in the multi-field-as-determinable view.

particles follow the actual trajectories guided by the guiding equation, but even when these particles are at space-like distance, the determinate value of the multi-field at a given time will depend on the exact location of the particles at that time. This is the way in which the multi-field accommodates the experimental violation of Bell's inequalities: the determinate cannot be locally defined, its value will be defined at any instant only by the actual configuration of the Bohmian particles, independently from the distance between the particles.

For example, given a 2-particle entangled state:

$$\psi(x_1, x_2) = c_1\psi_1(x_1)\psi_2(x_2) + c_2\psi_2(x_1)\psi_1(x_2) \qquad (6.1.1)$$

with actual configuration $q = (q_1, q_2)$, when the entangled state describes a macroscopic superposition, e.g. when the two components $\psi_1(x_1)\psi_2(x_2)$ and $\psi_2(x_1)\psi_1(x_2)$ are at a macroscopic distance with each other (this is also the case of space-like separated components) the Bohmian particles (q_1, q_2) will enter nly one of the two components, giving rise to the effective factorization.[36] As a result, we have two possible cases:

1. $\psi_1(q_1)\psi_2(q_2)$ with probability $P = |c_1|^2$ (6.1.2)
2. $\psi_2(q_1)\psi_1(q_2)$ with probability $P = |c_2|^2$ (6.1.3)

Repeating the experiment several times, this will result in the usual non-local correlations described by Bell's theorem. Note that every time the multi-field will have a determinate value described by $\psi_1(q_1)\psi_2(q_2)$ or $\psi_2(q_1)\psi_1(q_2)$.

[36] The process of effective factorization or *effective collapse* has been originally introduced in Bohm & Hiley (1987). In short, the effective factorization is the process that originates *effective wave functions* from larger entangled states when the latter describe macroscopic superpositions. This is the Bohmian equivalent of the branching process in Many Worlds Interpretation. Note that the formation of effective wave functions (EWFs) is independent from the interaction with the measuring apparatus. For example: in Bohm's theory, the entanglement between the system and the external environment produces EWFs (see e.g. Romano 2023). The formation of EWFs is the physical basis of decoherence in Bohm's theory.

6.2. Multi-field and the guiding equation

It must be noticed that, even though the multi-field assigns indeterminate values to most points of the wave function, the velocity of the Bohmian particles, described by the guiding equation, is defined for the N-tuples of points corresponding to the actual location of the Bohmian particles. And for these points the multi-field assigns a determinate. For the empty points (corresponding to indeterminate values of the multi-field) the guiding equation can still be defined, but it does not correspond to a real velocity of the particles. In other words, the guiding equation defines a velocity field for all points of the wave function, but the actual velocity of the particles is defined only for the points occupied by the particles. For these points the multi-field has a determinate. This grounds an ontological correspondence between the multi-field as determinate and the real velocity of the particles. The particles' velocity is always defined at their actual location, and the actual location of the particles correspond to the N-tuple of points for which the multi-field assigns a *determinate*.

6.3. Determinate and indeterminate knowledge

From the ontological point of view, the multi-field assigns a unique determinate at any instant. The determinate is assigned at the N-tuple of points where the Bohmian particles are located. However, from the epistemic point of view, the exact position of the Bohmian particles is unknown and statistically distributed according to: $\rho(q) = |\psi(q,t)|^2$. Consequently, the maximum knowledge we can have of the determinate value of the multi-field will be also statistically distributed according to the Born's rule. The fact that we do not know epistemically the exact configuration at a given instant, however, is not relevant for the ontology of the multi-field: independently from our knowledge, the state of affair (metaphysically speaking) is determinate: there is a unique location of the particles at every instant, which corresponds to a unique determinate of the multi-field and many (potentially infinite) indeterminate values for the unoccupied points. To this regard, the multi-field does not pretend to explain why the Bohmian particles are statistically distributed according to the Born's rule, or why this statistical distribution represents an ultimate epistemic constraint. This is an assumption that we have to maintain in the multi-field account, as it happens in all other metaphysical interpretations of the wave function in Bohm's theory, such as the nomological and the realist interpretation in configuration space.

7. Conclusions

I proposed that the multi-field can be characterized in metaphysical terms as a determinable, as it assigns to each point of 3D space a set of possible, potentially infinite, complex values and a determinate to the N-tuples of points which correspond to the exact location of the Bohmian particles. The multi-field so defined is a case of "gappy" metaphysical indeterminacy: it describes an indeterminate state of affairs in which a determinable property is instantiated by a set of possible determinates. We also noted that the determinate specified by the multi-field is non-local, as it depends from the position of the Bohmian particles of the actual configuration. When regarded under this approach, the pilot-wave of the de Broglie--Bohm's theory becomes an object less concrete and more abstract than a classical wave, but one that guides physically the particles in 3D space.

Acknowledgments

I wish to thank the audience at the Triennial International Conference of the *Italian Society for Logic and the Philosophy of Science* in Urbino 2023, where this paper was originally presented. This work has been supported by the *Fundação para a Ciência e a Tecnologia* through the fellowship FCT Junior Researcher hosted by the Centre of Philosophy of the University of Lisbon.

References

Albert, D. Z. (2013), "Wave function realism", in D. Albert & A. Ney (eds.), *The Wave Function: Essays on the Metaphysics of Quantum Mechanics*, Oxford, Oxford University Press.

Bacciagaluppi, G. & Valentini, A. (2009), *Quantum Theory at the Crossroads: Reconsidering the 1927 Solvay Conference*, Cambridge University Press.

Ballentine, L. E. (2008), "Classicality without decoherence: a reply to Schlosshauer", in *Foundations of Physics*, n. 38, pp. 916-922.

Ballentine, L. E. (2014), *Quantum Mechanics: A Modern Development*, World Scientific Publishing.

Barnes, E. (2010), "Ontic vagueness: a guide for the perplexed", in *Noûs*, vol. 44, n. 4, pp. 601-627.

Barnes, E. & Williams, R. (2011), "A theory of metaphysical indeterminacy", in K. Bennett & D. W. Zimmerman (eds.), *Oxford Studies in Metaphysics*, vol. 6, Oxford, Oxford University, pp. 103-148.

Belot, G. (2012), "Quantum states for primitive ontologists: a case study", in *European Journal for Philosophy of Science*, n. *2*, pp. 67-83.

Bohm, D. (1952), "A suggested interpretation of the quantum theory in terms of "hidden" variables", part I and II, in *Physical Review*, vol. 85, n. 2, pp. 166–193.

Bohm, D. & Hiley, B. J. (1987), "An ontological basis for the quantum theory: non-relativistic particle systems", in *Physics Reports* vol. 144, n. 6, pp. 321-375.

Bohm, D. & Hiley, B. J. (1993), *The Undivided Universe: An Ontological Interpretation of Quantum Theory*, Routledge.

Bokulich, A. (2014), "Metaphysical indeterminacy, properties, and quantum theory", in *Res Philosophica*, vol. 91, n. 3, pp. 449-475.

Bowman, G. E. (2008), *Essential Quantum Mechanics*, Oxford, Oxford University Press.

Callender, C. (2015*),* "One world, one beable", in *Synthese*, vol.192, n. 10, pp. 3153-3177.

Calosi, C. & Mariani, C. (2021), "Quantum indeterminacy", in *Philosophy Compass*, vol. 16, n. 4, e12731.

Calosi, C. & Wilson, J. (2018), "Quantum metaphysical indeterminacy", in *Philosophical Studies*, pp. 176, 1-29.

Calosi, C. (2021), "Gappy, glutty, glappy", in *Synthese*, *vol. 199*, n. 3, pp. 11305-11321.

Fletcher, S. C. & Taylor, D. E. (2024), "The representation and determinable structure of quantum properties", in *Synthese*, vol. 204, n. 44.

Forrest, P. (1988), *Quantum Metaphysics*, Blackwell.

Ghirardi, G. C., Rimini, A. & Weber, T. (1986), "Unified dynamics for microscopic and macroscopic systems", in *Physical review D*, vol. *34*, n. 2: 470.

Goldstein, S. & Zanghì, N. (2013), "Reality and role of the wave function in quantum theory", in D. Albert & A. Ney (eds.), *The Wave Function: Essays on the Metaphysics of Quantum Mechanics*, Oxford, Oxford University Press.

Hubert, M. & Romano, D. (2018), "The wave function as a multi-field", in *European Journal for the Philosophy of Science*, vol. 8, pp. 521-537.

Joos, E. Zeh, H. D., Kiefer, C., Giulini, D. J., Kupsch, J. & Stamatescu, I. O. (2013), *Decoherence and the Appearance of a Classical World in Quantum Theory*, Berlin, Springer.

Landau, L. D. & Lifshitz, E. M. (2013), *Quantum Mechanics: Non-Relativistic Theory*, vol. 3, Elsevier.

Maudlin, T. (2019), *Philosophy of Physics: Quantum Theory*, Princeton, Princeton University Press.

Ney, A. (2021), *The World in the Wave Function: A Metaphysics for Quantum Physics*, Oxford, Oxford University Press.

Norsen, T. (2017), *Foundations of Quantum Mechanics*, Berlin, Springer.

Romano, D. (2021a), "Multi-field and Bohm's theory", in *Synthese*, vol. 198, n. 11, pp. 10587-10609.

Romano, D. (2021b), "On the alleged extra-structures of quantum mechanics", in *Foundations of Physics*, vol. 51, n. 1, pp. 1-19.

Romano, D. (2022), "The unreasonable effectiveness of decoherence", in V. Allori (ed.): *Quantum Mechanics and Fundamentality*, *Naturalizing Quantum Theory between Scientific Realism and Ontological Indeterminacy*, vol. 460, Springer, pp. 3-18.

Romano, D. (2023), "A decoherence-based approach to the classical limit in Bohm's theory", in *Foundations of Physics*, vol. 53, n. 41, part of the special issue by A. Drezet (ed.): *Pilot-wave and beyond: Louis de Broglie and David Bohm's quest for a quantum ontology*, Foundations of Physics, 2023.

Rovelli, C. (1996), "Relational quantum mechanics", in *International Journal of Theoretical Physics*, vol. 35, pp. 1637-1678.

Sakurai, J. J. & Napolitano, J. (2020), *Modern Quantum Mechanics*, Cambridge, Cambridge University Press.

Schlosshauer, M. A. (2007), *Decoherence and the Quantum-to-Classical Transition*, Berlin, Springer.

Schlosshauer, M. A. (2019), "Quantum decoherence", in *Physics Reports*, vol. 831, pp. 1-57.

Shankar, R. (2012), *Principles of Quantum Mechanics*, Springer Science & Business Media.

Wallace, D. (2010), "Decoherence and ontology, or: how I learned to stop worrying and love FAPP", in S. Saunders, J. Barrett, A. Kent & D. Wallace (eds.), *Many Worlds? Everett, Quantum Theory and Reality*, Oxford, Oxford University Press.

Wallace, D. (2012), *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*, Oxford, Oxford University Press.

Wilson, J. (2013), "A determinable-based account of metaphysical indeterminacy", in *Inquiry*, vol. 56, n. 4, pp. 359-385.

Wilson, J. (2017), "Determinables and determinates", *Stanford Encyclopedia of Philosophy* (first published: 2017; substantive revision: 2023). Online:

https://plato.stanford.edu/archives/spr2023/entries/determinate-determinables/
Zurek, W. H. (2002), "Decoherence and the transition from quantum to classical - revisited", in *Los Alamos Science*, vol. 27, pp. 86-109.

The volume showcases a selection of articles presented at the Triennial Conference of the SILFS (Italian Association for Logic and Philosophy of Sciences) held at the University of Urbino on 4-7 September 2023. The authors are established scholars or junior researchers based both in Italian and international academic institutions. The articles featured in the volume afford a variety of themes and points of view on long-standing, as well as on others recently arisen, issues in the philosophy of sciences spanning, in particular, such areas as philosophy of physics, philosophy of mathematics, philosophy of cognitive sciences and AI, and general philosophy of science. By combining technical sophistication with breadth of philosophical analysis, these works provide readers with a distinct sense of progress and direction in fundamental debates in the philosophy of sciences.

ISSN 2037-4348

The volume showcases a selection of articles presented at the Triennial Conference of the SILFS (Italian Association for Logic and Philosophy of Sciences) held at the University of Urbino on 4-7 September 2023. The authors are established scholars or junior researchers based both in Italian and international academic institutions. The articles featured in the volume afford a variety of themes and points of view on long-standing, as well as on others recently arisen, issues in the philosophy of sciences spanning, in particular, such areas as philosophy of physics, philosophy of mathematics, philosophy of cognitive sciences and AI, and general philosophy of science. By combining technical sophistication with breadth of philosophical analysis, these works provide readers with a distinct sense of progress and direction in fundamental debates in the philosophy of sciences.

ISSN 2037-4348

Claudio Ternullo, Matteo Antonelli, ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

ARTIFICIAL MINDS, REALISM AND EVIDENCE IN SCIENCE

Proceedings of the 2023 Triennial Conference of the Italian Association for Logic and Philosophy of Sciences (SILFS)

edited by

Claudio Ternullo, Matteo Antonelli

Isonomia *Epistemologica*