# Mathematics, Language Games, and Black Boxes from Galileo to Wittgenstein

**Giovanni Galli**

## 1. Introduction

With the advent of *Large Language Models* (LLMs) like ChatGPT and Bard, there has been a surge of interest in whether these systems can truly understand the words they use, akin to human skills. This curiosity extends to whether such models can replicate human-like cognitive abilities. The development of LLMs also prompts more profound questions about the relationship between language and mathematics, particularly concerning how meaning can be encoded algorithmically. While the debate over machine understanding of language is not new, it is gained renewed importance as AI systems grow with a more sophisticated architecture. Some developers now claim that these systems exhibit a form of understanding, a perspective that merits closer philosophical scrutiny. Mathematics plays a pivotal role in the architecture of LLMs, embedding foundational assumptions within these models. This paper argues that those who advocate for LLMs' capacity to understand meaning implicitly adopt a Galilean view of mathematics. This view is at odds with the later Wittgenstein's philosophy, which posits that mathematics is not merely a reflection of the world's ontological structure but is fundamentally a human activity. The shift from Galileo's to Wittgenstein's views on mathematics provides crucial insights into the intersections of mathematics, language, and the limits of AI. These insights are especially relevant when considering Wittgenstein's private language argument (PLA), which challenges the possibility of a private language and has significant im-


DOI 10.35948/DILEF/978-88-6032-827-4.04

**Giovanni Galli**, University of Teramo, ggalli@unite.it, 0000-0002-8289-345X

plications for the development of NLP systems. In particular, the problem addressed here is that of the LLMs' ability to understand and create meanings with the same human-like features. The discussion about the private language pivots this issue insofar as it is a case limit for human understanding and meaning formation. The underlying assumptions is that if a LLM is able to model private language, then it understands and uses meaningfully words in language. The analysis I propose is that we can distinguish two mathematical frameworks grounding the LLMs' ability to understand meaning and model private language, namely the Galilean and the Wittgensteinian. The Galilean framework is implied in defending the ability of LLMs to understand, while the Wittgensteinian accommodates the claim that the LLMs ability to manipulate meaning are different from human practices. Wittgenstein's ideas, especially his concept of language games, are foundational in the design of many natural language processing (NLP) systems and LLMs[1]. In recent years, these systems have achieved unprecedented levels of performance in tasks such as question answering, textual entailment, and machine translation, utilizing deep learning neural networks (DLNNs) to learn from vast amounts of data. Examples of these advancements include works by Devlin *et al*. (2019), Kitaev, Cao, and Klein (2019), and Wang et al. (2019)[2].

1 Stephen Mills, *Wittgenstein and Connectionism: A Significant Complementarity?*, in «Royal Institute of Philosophy Supplement», 34, 1993, pp. 137-157; Charles W. Lowney, Simon D. Levy, William Meroney, Ross W. Gayler, *Connecting Twenty-First Century Connectionism and Wittgenstein*, in «Philosophia», 48, 2020, pp. 643-671; Ines Skelac, Andrej Jandrić, *Meaning as Use: From Wittgenstein to Google's Word2vec*, in Sandro Skansi (ed.), *Guide to Deep Learning Basics. Logical, Historical and Philosophical Perspective*, Berlin, Springer, 2020, pp. 41-53.

2 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, Association for Computational Linguistics, 2019; Nikita Kitaev, Steven Cao, and Dan Klein, *Multilingual Constituency Parsing with Self-Attention and Pre-Training*, In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3499–3505, Florence, Italy, Association for Computational Linguistics, 2019;

This paper examines Wittgensteinian concepts that influence the architecture of many deep learning NLP systems, with a particular focus on efforts to model a private language. By dissecting these concepts, I aim to uncover the assumptions that underpin AI systems' language modeling capabilities. Specifically, I will explore key features of NLP systems[3] used for word embedding and evaluate a proposal by Lowney *et al*. (2020) to manipulate a form of private language through a vector symbolic architecture.

In the following section §2, I will compare the fundamental issues of Galilei and Wittgenstein's philosophy of mathematics. In §3, the private language argument (PLA) is described. In §4, I will discuss connectionist language models and introduce some notions about NPL systems' architecture on the light of Wittgenstein's and Galilei's views about mathematics. An overview of this model helps introduce the work of Lowney, Levy, Meroney, and Gayler (2020). They submit that their model can respond to the issues raised by Wittgenstein in the notorious Beetle in the box case, used to illustrate Private Language Argument (PLA). This argument unexpectedly turned out to be relevant not only for the philosophy of language but also for NLP modelers. I will describe the language game concept in NLP, how it is embedded, and its role in inductive systems development. This central concept in Wittgenstein's work is relevant to describing context's role in understanding word meanings. In §5, I criticize Lowney and colleagues' claim, whose model does not successfully capture Wittgenstein's Beetle in the Box case. Moreover, I argue that even if we can distinguish

---

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman, *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*, Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, 2019, Article 294, 3266–3280.

3 For a discussion about language acquisition models, instead, see Jordi Poveda, Alfredo Vellido, *Neural Network Models for Language Acquisition: A Brief Survey*, in Emilio Corchado, Hujun Yin, Vicente Botti, Colin Fyfe (eds.), *IDEAL Intelligent Data Engineering and Automated Learning*, Berlin, Springer, 2006. In these pages, the focus will be on the only NLP models.

a strong and a weak definition of a private language, Wittgenstein's argument also holds for deep-learning models, and his worries are still a good guide for NLP and LLMs developers.

## 2. Galilei and Wittgenstein's Philosophy of Mathematics

The 400th anniversary of Galileo Galilei's *The Assayer* provides an opportunity to revisit its philosophical significance, particularly in relation to mathematics and scientific methodology. The Galilean dialogic style expresses a new, cutting-edge conception of mathematics and epistemology. His work presents a realist view of mathematics[4], where mathematical structures are seen as intrinsic to the physical world. In contrast, Wittgenstein[5], especially in his later work, adopts an anti-

4 There is no space here to dig into the intricacies of Galilei and Wittgenstein's thoughts about mathematics. For a more detailed study about the former, see Paolo Galluzzi, *Il "platonismo" del tardo Cinquecento e la filosofia di Galileo*, in Paola Zambelli (ed.), *Ricerche sulla cultura dell'Italia moderna*, Roma-Bari, Laterza, 1973, pp. 39-79; Mark A. Peterson, *Galileo's new mathematical philosophy*, in «Forum Italicum», 49(1), 2015, pp. 119-138; Carla Rita Palmerino, *Reading the book of nature: the ontological and epistemological underpinnings of Galileo's mathematica realism*, in Geoffrey Gorham, Benjamin Hill, Edward Slowik (eds.), *The language of nature: reassessing the mathematization of natural philosophy in the seventeenth century*, University of Minnesota Press, Minneapolis, 2016, pp 29-50.

5 We can distinguish different phases in the development of the philosophy of mathematics in Wittgenstein's work. I will only focus on the later Wittgenstein in these pages. For an in-depth analysis on Wittgenstein's philosophy of mathematics, see Crispin Wright, *Wittgenstein on the Foundations of Mathematics*, Londom, Duckworth, 1980; Michael Wrigley, *Wittgenstein on Inconsistency*, in «Philosophy», 55(214), 1980, pp. 471-484; Carlo Penco, *Matematica e gioco linguistico. Wittgenstein e la filosofia della matematica del* '900, Firenze, La Nuova Scuola, 1981; Michael Dummett, *Reckonings: Wittgenstein on Mathematics*, in «Encounter», 50(3), 1978, pp. 63-68; Jacques Bouveresse, *Le pays des possibles, Wittgenstein, les mathématiques et le monde réel*, Paris, Les Editions de Minuit, 1988; Diego Marconi, *Wittgenstein on Contradiction and the Philosophy of Paraconsistent Logic*, in «History of Philosophy Quarterly», 1(3), 1984, pp. 333-352; Pasquale Frascolla, *Wittgenstein's Philosophy of Mathematics*, London and New York, Routledge, 1994; Christine Redecker,

realist stance, viewing mathematics as a product of human activity rather than as a discovery of pre-existing truths. The comparison between Galilean realism and Wittgensteinian antirealism about mathematics could lead to an intriguing debate about the objects, nature, and specific features of mathematics. However, we must distinguish the differences in the particular contexts and philosophical purpose. Indeed, Wittgenstein's thoughts about mathematics were essential in determining his view about the meaning and the understanding of language, built on the notion of the "language game." In contrast, Galileo's view and use of mathematics had a fundamental impact on the future development of science. Galilei's view of mathematics is bound to a new conception of matter, which can be studied by measuring its primary qualities while discharging the secondary quality to the subjective and private sphere of experience. The view of mathematics proposed by Wittgenstein is closely tied to his beliefs about language. According to Galileo, language boundaries are defined by private objects, such as the secondary qualities of things. Therefore, I suggest that the connection between Galileo's and Wittgenstein's ideas about mathematics and the meaning of private objects is a significant topic that has yet to be thoroughly investigated.

The representation of the universe as a whole «written in a mathematical language» was not merely a metaphor to emphasize the epistemic role of mathematics. Still, it concerns Galilei's conception of maths as revealing the ontological fabric of reality:

---

*Wittgensteins Philosophie der Mathematik: Eine Neubewertung im Ausgang von der Kritik an Cantors Beweis der Überabzählbarkeit der reeleen Zahlen* [Wittgenstein's Philosophy of Mathematics: A Reassessment Starting From the Critique of Cantor's Proof of the Uncountability of the Real Numbers], Frankfurt-Hausenstamm, Ontos Verlag, 2006; Mathieu Marion, Mitsuhiro Okada, *Wittgenstein et le lien entre la signification d'un énoncé mathématique et sa preuve*, in «Philosophiques» 39 (1), 2012, pp. 101-124; Michael Potter, *Reason's Nearest Kin: Philosophies of Arithmetic from Kant to Carnap*, Oxford, Oxford University Press, 2000; Simo Säätelä, *From logical method to 'messing about': Wittgenstein on 'open problems' in mathematics*, in Oskari Kuusela, Marie McGinn (eds.), *The Oxford Handbook of Wittgenstein*, Oxford, Oxford University Press, 2011.

> La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi agli occhi (io dico l'universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile intenderne umana parola; senza questi è un aggirarsi vanamente per un oscuro laberinto[6].

According to Galilei, through mathematics, scientists gained the ability to capture the fundamental proprieties of natural phenomena to unveil the universe's structure underlying the subjective perceptions generated by our senses[7]. The image of mathematics as a language is a description that will travel the time from Galilei to Wittgenstein, shifting from the realist idea of a distinctive mathematical ontology of the universe to the anti-realist contemporary idea that it is a result of human activity. Given the ontological feature of mathematics, characterizing the objective structure of the universe, Galilei claims that we should distinguish objective from subjective properties of objects. This distinction discriminates between the qualities behold to the physical objects and the ones that reside in the perceiving subjects[8]. The relation between the secondary qualities of objects and the world is still much debated. This issue raises a seduction that connects many crucial turning points in the history of philosophy. It begins with Gorgia's nihilism about the possible connection between language and the world. It con-

6 I propose this translation: «Philosophy is written in this grand book, the universe, which stands continually open to our gaze. But the book cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics, and its characters are triangles, circles, and other geometric figures, without which it is humanly impossible to understand a single word of it; without these, one is wandering about in a dark labyrinth» (Galileo Galilei, *Il Saggiatore*, Michele Camerota, Franco Giudice (eds.), Milano, Hoepli, 2023, pp. 46-47).

7 Michele Camerota, Franco Giudice, *Introduzione. «La strada al ritrovamento del vero», Il Saggiatore come manifesto del nuovo sapere*, in *Il Saggiatore,* Michele Camerota, Franco Giudice (eds.), Milano, Hoepli, 2023, pp. xix.

8 Galileo Galilei, *Il Saggiatore*, cit., pp. 246-247.

tinues with Galilei's discussion about the secondary quality of objects in *The Assayer* (1623). Moreover, of course, it also touches on Wittgenstein's work. According to Galilei[9], the doctrine of secondary qualities (SQs) differs from primary accidents in size, shape, motion, rest, and location. Galileo claimed that primary accidents exist in external bodies, which we attribute to them. We give public ontology to the primary accidents by using a public language. On the other hand, we assign a private ontology to the secondary quality since these qualities «appear to exist in the objects we perceive around us and reside only in us»[10]. If the SQs reside exclusively in us, they can be considered private objects in a Wittgensteinian sense. Wittgenstein critiques the concept of a private ontology, offering an alternative perspective on how language operates about ourselves and the world. In PI §256 and §258–60, Wittgenstein claims that a private language is impossible in which meanings are derived by ostensive internal relation to private objects. The conclusion is that a meaningful discussion about private sensation is impossible.

In his preface to the Italian edition of Wittgenstein's *On Certainty*, Aldo Gargani writes:

> Sembra giusto ritenere che la scienza, soprattutto a partire dal secolo XVII, con la sua strutturazione meccanicistica, abbia scisso il sapere dal senso comune. La tesi di Galilei, Descartes, Hobbes e di altri sulla soggettività delle qualità sensibili ha espropriato dall'universo fisico oggettivo sapori, odori, colori (e insieme anche valori etici ed estetici) con i quali il senso comune produce la sua percezione del mondo fisico. La scienza avrebbe così svalutato i canali ordinari attraverso i quali il senso comune stabilisce il proprio contatto con gli oggetti fisici[11].

9 See Galileo Galilei, *Two Kinds of Properties. Selection from Il Saggiatore*, in *Philosophy of Science*, Arthur C. Danto, Sidney Morgenbesser (eds.), New York, Meridian Books, [1623] 1960, p. 28.

10 Eugen Fisher, *Philosophical Pictures and Secondary Qualities*, in «Synthese», 171(1), pp. 77-110: 93.

11 Aldo Gargani, *Scienza, filosofia e senso comune*, in Ludwing Wittgenstein, *Della Certezza*, Aldo Gargani (ed), Mario Trinchero, (translation), Torino, Einaudi, 1999, p. vii.

The Galilean view implies a twist of the traditional conception of sensitive qualities. The expressions of individual sensations seem to imply privacy of language; this is the «primitive seduction of the private language»[12]. These qualities, which Galileo relegated to the private sphere, align with Wittgenstein's critique of private language, i.e. the idea that meanings derived from internal, private experiences are incommunicable. The discussion about private sensations then found its acme in Wittgenstein's treatment of the PLA, which explores this idea further, asserting that a language grounded in private sensations is inherently flawed because such sensations cannot be meaningfully expressed or understood by others. This argument, central to Wittgenstein's later philosophy, has significant implications for AI, particularly in the development of LLMs that aim to model human language. It also has a vital role in developing the reflection about the nature and features of mathematical propositions and objects. According to Wittgenstein, «the mathematician is not a discoverer: he is an inventor» (RFM, Appendix II, par 2; LFM 22, 82). Anything exists mathematically before we humans have invented it. In the context of LLMs, the contrast between Galileo and Wittgenstein's views highlights the limitations of current AI models. While LLMs can process and generate text based on statistical patterns in data, they lack the ability to grasp the contextual nuances that are central to human language. This limitation is closely tied to Wittgenstein's critique of private language: just as a private language is impossible and cannot convey meaning to others, LLMs struggle to model the context-dependent understanding that humans naturally possess.

Moreover, Wittgenstein's notion of language games emphasizes that meaning is not fixed but is shaped by use within specific contexts. This idea challenges the premise that LLMs can achieve genuine understanding, as these models operate primarily through pattern recognition rather than through an appreciation of the contextual factors that give language its meaning. Consequently, while LLMs may simulate understanding, they do not possess the human-like capability to fully

12 Robert J. Fogelin, *Wittgenstein*, London and Boston, Routledge, 1976, pp. 156.

engage with the complexities of language. The attempt by Lowney *et al.* (2020) to model a private language using neural networks is particularly illustrative of these challenges. Despite their claims, this paper argues that their model fails to capture the essence of Wittgenstein's beetle in the box analogy, which illustrates the impossibility of private language. Wittgenstein's argument remains relevant, suggesting that the challenges he identified continue to pose significant obstacles for AI developers.

In conclusion, this paper argues that the mathematical underpinnings of LLMs, rooted in a Galilean view of mathematics, are inadequate for capturing the full richness of human language. Wittgenstein's insights into the nature of language and meaning offer a crucial counterpoint, highlighting the limitations of current AI approaches and suggesting that understanding remains beyond the reach of machines, as long as they lack the ability to engage with language in a genuinely contextual and human way.

## 3. The Private Language Argument

In this section, I present the PLA and characterize it as a language game. The PLA is one of the most famous contributions of the later Wittgenstein. According to Fogelin (1976, 153), PLA is Wittgenstein's most debated argument. The debate sparked by PLA, already broad in the 70s, even increased over subsequent years and is still one of the most discussed aspects of Wittgenstein's philosophy to date.

Furthermore, in *Philosophical Investigations* (PI), Wittgenstein offers two main scenarios[13]: the first is the example of the diary of the oc-

13 For a critical analysis about private language argument and LLMs, see Giovanni Galli, *Language Models and the Private Language Argument: a Wittgensteinian Guide to Machine Learning*, in Brian Ball, Alice Helliwell, Alessandro Rossi (eds.), *Wittgenstein and Artificial Intelligence, Vol. I, Mind and Language*, London, Anthem Press, 2024, pp. 145-163.

currence of a private sensation (PI, §258), and the other is the well-known case of the beetle in the box (PI, §293). The argument sketch is presented in PI §243: «The words of this [private] language are to refer to what only the speaker can know – to his immediate private sensations. So another person cannot understand the language». His attack on the idea and the possibility of a private language is contained in the passages of PI §§244- 271. It is not here the place to disentangle the exegetical complexity and the interpretative genealogy of the PLA. It is highly controversial whether there is, or there at least could be, a specific private language argument (PLA) to be found in Wittgenstein. However, it will be enough to cite the four central positions: orthodox, Kripkean, substantial, and resolute. While the orthodox way[14] claims that the PLA is a reductio ad absurdum argument, Kripke[15] argues that the PLA is the consequence of the discussion on rule-following. In general, the substantial view argues that private language is impossible. In contrast, according to Mulhall (2007, 18), language limitation is simply nonsense: no sense can be given to the concept of a philosophically substantial private language.

Hacker states that the PLA is one of «he most original and significant philosophical reflections of the twentieth century. If the line of argument pursued in them is valid, their implications, both within philosophy and without, are considerable. Modern philosophical logic, theoretical linguistics, and empirical psychology branches would need

14 Norman Malcom is one of the founders of the PLA (see Norman Malcom, *Wittgenstein's Philosophical Investigations*, in «Philosophical Review», 63(4), 1954, pp. 530-559) and one of the defenders of what lately was called the *orthodox view* of PLA. That is the idea that the paragraphs §§244-271 contain embedded an argument in the form of *reductio ad absurdum*.

15 Kripke argued that Wittgenstein introduced a new skeptical problem to which he gave a Humean solution. According to Kripke, PLA is connected to the logical and epistemological character of following a rule. I agree instead with Hacker's interpretation of the PLA (G. P. Backer, P. M. S. Hacker, *On misunderstanding Wittgenstein: Kripke's private language argument*, in «Synthese», 58(3), 1984, pp. 407-450; Peter M. S. Hacker, *Wittgenstein: Connections and Controversies*, Oxford, Clarendon Press, 2001).

re-evaluation»[16]. Moreover, perhaps we can also add branches of artificial intelligence today. Lowney, Levy, Meroney, and Gayler (2020) intentionally selected the beetle in the box as a case study to demonstrate the integration of contextual information into a language model, even if it is sourced from a private ontology. The "beetle in the box" case describes a scenario in which everyone has a box with something in it. This "something" is called a "beetle". Anyone can look inside each box, and everybody asserts that they know what a "beetle" is, only looking at his "something-beetle" in the box. The thing in the box is unlikely to be the same for all. Indeed, it could be that the box is empty. The point is that «the thing in the box has no place in the language game. […] That is to say: if we construe the grammar of the expression of sensation on the model of 'object and designation', the object drops out of consideration as irrelevant» (PI, §293). The private "something" to which Wittgenstein refers is not a simple invention. It is an example of a private experience. It could be a private sensation, definition, or object[17]. I submit that §293 of Wittgenstein's works can be interpreted as a scenario defining the limit for language games. Wittgenstein describes language as played through language games, and it is possible to argue that there is a specific limit to language and language games. The PLA identifies this limit precisely with the beetle in the box case. The "thing in the box", or perhaps the same box being empty, falls outside the boundaries of the context where we can find the meaning of the words we use.

Wittgenstein's dictum about language games, meaning use, family resemblance, context, and the tenets of connectionism all have similarities, and the powerful and flexible concepts discussed in PI lead to PLA being the primary hotspot of the discussion. Therefore, using PLA to understand the connection between Wittgenstein's work and connectionism is more relevant than succumbing to simplistic interpretations. In the upcoming section, I will examine the connection between

16 Peter M. S. Hacker, *Wittgenstein: Connections and Controversies*, Oxford, Clarendon Press, 2001, p. 209.

17 Candlish Stewart, *Private Objects and Experimental Psychology*, in Annalisa Coliva, Eva Picardi (eds.), *Wittgenstein Today*, Padova, Poligrafo, 2004, pp. 297-317.

Galilei's and Wittgenstein's views and connectionism. I will specifically concentrate on two major NLP models, one used for word embedding and the other used to simulate the beetle-in-the-box scenario.

## 4. The Connectionist Framework of Machine-Learning Language Models

The shift from symbolic AI to connectionist models can be seen as a contemporary echo of the transition from Aristotelian to Galilean science. Galileo's rejection of purely qualitative descriptions in favour of quantitative, mathematical analysis parallels the move from rule-based symbolic systems to the data-driven, mathematically grounded approach of connectionism. Connectionist frameworks, which rely on distributed representations and pattern recognition, resonate with Galileo's emphasis on observation and mathematical description. Just as Galileo's mathematical formulations provided a new way to understand the natural world, connectionist models offer a new paradigm for understanding cognitive processes and language: «connectionist models provide a new paradigm for understanding how information might be represented in the brain»[18]. Both approaches highlight the importance of mathematics as a tool for revealing underlying structures, whether in the physical world or in the domain of language and cognition. In this context, Wittgenstein's emphasis on the use of language in context aligns with Galileo's approach to mathematics as a practical tool rather than an abstract, isolated system. The success of connectionist models in natural language processing is thus not just a triumph of computational power but also a reflection of the enduring power of mathematical descriptions to model and make sense of complex systems. By integrating Galileo's philosophy of mathematics into our understanding of propositions and connectionist models, we can appreciate how these frameworks continue the Galilean tradition of using mathematical structures to describe

18 Cameron Buckner, James Garson, *Connectionism*, in Edward N. Zalta (ed.), «The Stanford Encyclopedia of Philosophy» (Fall 2019 Edition), URL = https://plato.stanford.edu/archives/fall2019/entries/connectionism/.

and understand the world. Whether in the realm of physical laws or linguistic patterns, the influence of Galileo's approach is evident in the ongoing evolution of how we model and understand reality.

### 4.1. What is a proposition?

What is a proposition? This is the central question that TLP, PI, and NLP connectionist methodologies aim to answer. Wittgenstein writes in the *Notebook*: «My whole task consists in explaining the nature of the proposition» (T, 22.1.15). From this line, we must begin to trace back to the work of Wittgenstein to enlighten the contemporary exercise of connectionist methodologies for natural language. Wittgenstein's main objective was to explain the nature of propositions in his work. However, it is essential to note that the nature or essence of a proposition should not be conceived as a platonic feature of an item existing in a detached realm of beings, happenings, and words and statements we express. Instead, the nature of a proposition is interwoven in the use of words in our language. This is why the first connection point between Wittgenstein's philosophy and connectionism can be found in Notebook and TPL, even before PI. Wittgenstein's exploration of the nature of propositions aligns with the methodological rigor seen in Galileo's approach to mathematics. For Galileo, mathematics was the language of the universe, and its propositions were not just abstract entities but reflections of physical reality. This perspective suggests that the nature of a proposition, while deeply rooted in linguistic usage, also has a structural integrity that mirrors the mathematical order of the world. In the same way that Galileo saw mathematical truths as inherent in the fabric of reality, Wittgenstein viewed propositions as inherently tied to the structure of language and its usage within the world. The connection between Wittgenstein's ideas and contemporary connectionist methodologies can be further understood through this Galilean lens. Just as Galileo redefined the understanding of natural phenomena through mathematical propositions, connectionist models like Word2Vec redefine language understanding through the mathematical embedding of words. This redefinition echoes Galileo's mathematical realism, where propositions are seen as tools to decode and describe the world's order.

### 4.2. Word2Vec and the Philosophy of Language

The first connectionist NLP method to be studied under Wittgensteinian light is Word2Vec[19]. It is a group of models based on neural network systems that produce word embedding. It is one of the first machine learning methods used to represent words as vectors, but it now seems outdated due to the development of transformer models, i.e., ChatGPT and Bard.

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

$$f : x_i \to p_i$$

The input elements are $\{x_i\}_{i \in N} \subset \Omega_X$; and the output elements are *y-targets* $y_i \in \Omega_Y$. Usually, to parametrize $f$ is used a set of parameters which establishes a family of estimators for the given estimation. Training an ML estimator means to optimize the estimator using the parameter values, in order to fit the data at best according to a prescribed loss function. The process to find the optimal estimator for a given training set is to train the model.

The Word2Vec models produce word embedding, in which semantic structures, such as words, phrases, or similar entities from a specific vocabulary, are mapped to and mathematically modelled as Euclidean vectors of real numbers. It has a variety of applications,

19 One of the leading researchers which implemented firstly Word2Vec is Tomáš Mikolov, who introduced this technique in NLP in Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, 5 May 2022, https://arxiv.org/abs/1301.3781.

and it is helpful to generate text similarity, sentiment analysis, and recommendation systems. The system deploys vectorial distribution to assign a specific value to a word analysed in a context, a particular corpus. It will be likely to find in the vectorial space the word "cat" near "dog, pet, kitty, purr, paws, meow," with a value far distant from a word that could be defined as an alcoholic drink, which establishes the surrounding of, say, "wine" and "beer." Word2Vec can utilize either model architecture to produce a distributed representation of words. The representation of words defines the collocation of words and their interlinguistic connections. The two models in play are continuous bag-of-words (CBOW) and continuous Skip-gram. The CBOW model predicts the current word from a window of surrounding context words. The order of context words does not influence the prediction, which is the bag-of-words assumption. The Skip-gram model is the reverse. It uses the current word to predict the surrounding window of context words[20].

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

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

20 *Ibidem.*

["The," "beetle," "is"], the model will predict the probability of "The" and "is" before and after the target word "beetle."

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

Word2Vec, as a mathematical model for word representation, reflects a Galilean approach to language: it seeks to quantify and map linguistic phenomena using mathematical structures. Galileo's philosophy emphasized that the universe is written in the language of mathematics, and Word2Vec embodies this idea by representing words as vectors in a mathematical space. This method parallels how Galileo used mathematical propositions to describe physical laws, suggesting that language, too, can be understood through a mathematical framework.

However, just as Galileo's mathematical descriptions were limited by the observational tools of his time, Word2Vec's capacity to fully capture meaning is constrained by its reliance on collocations and context within a given corpus. Galileo acknowledged the limitations of his tools, recognizing that mathematical descriptions are approximations of a more complex reality. Similarly, Word2Vec's vectors are approximations, limited by the data they are trained on, and cannot fully encapsulate the multifaceted nature of language.

### 4.3. The Vector Symbolic Architecture

The second model to be scrutinized is the VSA, which Lowney and colleagues (2020) used to model the beetle-in-the-box case. VSA stands for Vector Symbolic Architecture, a connectionist model using high-dimensional vectors to encode systematic and compositional information as distributed representations[21]. VSA family of models follows the connectionist framework of Smolensky[22], extending it into high-dimensional vector space. Lowney, Levy, Meroney, and Gayler set up a formalism comprising three operations on vectors[23]: multiplication, addiction, and permutation. According to them, «VSA provides a principled connectionist alternative to classical symbolic systems (predicate calculus, graph theory) for encoding and manipulating various useful structures». They suggest that «The biggest advantage of VSA representations over other connectionist approaches is that a sin-

21 Pentti Kanerva, *The spatter code for encoding concepts at many levels*, in Maria Marinaro, Pietro Morasso (eds.), *ICANN '94, Proceedings of the international conference on artificial neural networks*, London, Springer-Verlag, 1994, pp. 226-229; Daniel Rasmussen, Chris Eliasmith, *A Neural Model of Rule Generalization in Inductive Reasoning*, in «Topics in Cognitive Science», 3, 2011, pp. 140-153.

22 Paul Smolensky, '*Connectionism, constituency and the language of thought*', In *Connectionism: Debates on psychological explanation* (Vol. 2), edited by MacDonald, Cynthia, and MacDonald, Graham, 164–198. Oxford: Blackwell, 1995; Paul Smolensky, '*Reply: Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture*', In Connectionism: Debates on psychological explanation (Vol. 2), edited by MacDonald, Cynthia, and MacDonald, Graham, 223–90. Oxford: Blackwell, 1995.

23 Charles W. Lowney II, Simon D. Levy, William Meroney, Ross W. Gayler, *Connecting Twenty-First Century Connectionism and Wittgenstein*, «Philosophia», 48, 2020, pp. 643-671.

gle association (or set of associations) can be quickly recovered from a set (or larger set) of associations in a time that is independent of the number of associations». In that way, «VSA thus answers the scalability problem raised by classicists about biologically plausible real-time processing»[24]. They chose this model to capture statements similar to the beetle case in PI. Their choice relies on «VSAs use multidimensional vectors and numerical weights, randomly assigned at the most basic level, in the actual processing of the networks constructed». The flexibility they attribute to the model is also based on the fact, which is the Wittgensteinian tenet against the ostensive relation to private objects, that «There is no one-to-one correspondence to an entity or item for representation. A symbol is represented in signs/vectors distributed across a vector space. Operations with symbols, in turn, use these distributed representations to establish proximity relations that model thought and language use»[25]. To recognize the meaning of a word as a symbol that does not have an ostensive and one-to-one «correspondence to an entity or item for representation» is specifically to rely on the idea that «for Wittgenstein there is not typically an atomic content or correspondence that one can point to explicate the meaning of a term»[26]. The meaning of a word, a symbol, is a product of «a complicated network of similarities overlapping and crisscrossing» (PI, §66;)[27]. Following Goldfarb (1997), Strawson (1954), and Hintikka & Hintikka (1986)[28], Lowney and the other researchers agree with the Hintikkas' way of thinking, who believe that «Wittgenstein was not de-

24 Ivi, p. 654.

25 *Ibidem*.

26 *Ibidem*.

27 See also Stephen Mills, *Wittgenstein and Connectionism: A Significant Complementarity?*, cit., p. 139.

28 Warren Goldfarb, *Wittgenstein on the Fixity of Meaning*, In *Early analytical philosophy: Frege, Russell, Wittgenstein. Essays in honor of Leonard Linsky*, edited by Tait, William, W., 75-89. Chicago: Open Court, 1997; Peter Frederick Strawson, *Wittgenstein's "Philosophical Investigations"*, Mind, 63, 70-99, 1954; Jaakko Hintikka, and Merrill B. Hintikka, Investigating Wittgenstein. Cambridge: Basil Blackwell, 1986.

nying the possibility of referring to sensations nor a private language outright»[29]. They use connectionism to shape a formalism in which Wittgensteinian assumptions about the nature of language are satisfied, namely that the VSA can capture some language features without assuming the connection with objects for meaning. However, the limitation of the beetle case remains fixed if we model it with a neural network model as VSA. Connectionist models cannot explain language and its meanings insofar as Wittgenstein stressed; we cannot give theories of language, only descriptions. Connectionist models, as mind models, can «help guide inquiry into the workings of the phenomena and can dispel some misconceptions, but as close as it may come to analogically portraying some important features, it should not be mistaken for the only or the actual way that language works»[30]. Perhaps it is better to say that the VSA proposed to model the beetle case does not resolve the beetle puzzle, even if it models the case following the line of the Hintikkas' and Hacker's interpretation, according to which there is no literal claim against the possibility of using the language to talk about private objects, as private sensations. Still, it is possible to talk about these private items using a language made by public meanings construed through interactions in extra-linguistic contexts. These contexts are not yet encoded in systems such as Word2Vec or the Smolensky vectors. With these corrections to Lowney and colleagues' proposal, I agree with their conclusion that «by respecting Wittgenstein's insights and providing a VSA account that displays linguistic compositionality, integrates soft symbols, and develops analogical structures that can be systematic and advance productively, we have shown how twenty-first-century connectionism can address what appeared to be limitations in the functionality of its operation, limitations in learning, and limitations in biological plausibility that might have thwarted connectionism's ability to be a better mind-model for language and cog-

29 Charles W. Lowney, Simon D. Levy, William Meroney, Ross W. Gayler, *Connecting Twenty-First Century Connectionism and Wittgenstein*, cit., p. 659.

30 Ivi, p. 668.

nitive science»[31]. In this section, I have presented Word2Vec and the VSA proposed by Lowney and colleagues (2020). The two systems have underlying philosophical assumptions that Wittgenstein developed. Consequently, they show how Wittgenstein's ideas are deeply embedded in the deep-learning NLP models and how his ideas are integral to the breakthrough of AI language models. The following section will explore how the PLA is relevant for the problem of black box in LLMs.

## 5. The Private Language Argument and Algorithmic Black Boxes

The black box problem in large language models (LLMs) highlights a fundamental challenge in AI and machine learning: the difficulty of understanding how these systems process and generate output from the input data. This issue becomes even more intricate when viewed through the lens of the philosophy of mathematics and language, particularly as explored by Wittgenstein and Galileo. On one hand, for Wittgenstein, the meaning of a term is not derived from an isolated, atomic content or direct correspondence with reality. Instead, meaning emerges from its use within a language game, which is a complex interplay of social interactions, context, and cultural norms. Wittgenstein's private language argument (PLA), exemplified by his "beetle in a box" thought experiment, asserts that private, internal experiences (like the sensation of pain or seeing a beetle) cannot be meaningfully communicated or even understood by others if there is no public criterion for their use. The beetle, locked away in a private box, is inaccessible to others and, by extension, meaningless within a shared, public, language framework. This idea parallels the black box problem in LLMs. Just as the beetle is hidden from view in Wittgenstein's analogy, the internal workings of a neural network – how it processes input data and arrives at its output – are often opaque and inaccessible to human understanding. Even the designers of these models might not fully grasp how specific pre-

31 *Ibidem*.

dictions or decisions are made, as Rudin and Radin[32] note: the algorithms combine variables in ways that are hidden from view, resulting in an output that is effectively a "black box." On the other hand, Galileo's contribution to this discourse lies in his mathematical treatment of the natural world. By abstracting phenomena into mathematical models, Galileo laid the groundwork for modern science's approach to understanding reality through quantitative analysis. However, while these models can predict and describe natural phenomena with precision, they do not necessarily grasp the underlying fundamental features of the phenomena themselves. Similarly, LLMs can generate human-like text and solve complex problems, but they do so without a human-like understanding of the content they process – much like Galileo's models, which describe but do not "understand" the world.

The black box problem comes from an overlapping of different issues: the opacity problem, the strangeness problem, the unpredictability problem, and the justification problem[33]. The opacity problem is articulated in different issues, including the representation learning of hidden layers. One influential discussion about it is carried out by Bengio *et al.* (2013)[34], which describes how DLNNs models must learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data. One example of the informative hidden layer representations is, as we have seen, the traditional word vectorization. Contemporary, more complex Transformer techniques, as Tamir and Shech note (2023: 337-38), use deeper pre-training of text embedding methods, which are adapted to embed chunks of text and individ-

32 Cynthia Rudin & Joanna Radin, *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition*, in «Harvard Data Science Review», 1 (2), 2019.

33 Bartosz Brożek, Michał Furman, Marek Jakubiec, Bartłomiej Kucharzyk, *The black box problem revisited. Real and imaginary challenges for automated legal decision making*, in «Artif Intell Law 32», pp. 427-440 (2024), https://doi.org/10.1007/s10506-023-09356-9.

34 Yoshua Bengio, Aaron Courville, and Pascal Vincent, *Representation Learning: A Review and New Perspectives*, IEEE Trans. Pattern Anal. Mach. Intell. 35, 8 (August 2013), 1798-1828. https://doi.org/10.1109/TPAMI.2013.50.

ual terms in the context of surrounding text in which they are written[35]. As the example of Word2Vec shows, Mikolov *et al*. (2013) use shallow neural networks to map indi«fill in the blank». According to Tamir and Shech (2023: 332), this vectorized word-embedding system is helpful for the original task. Still, the representations elaborated could be reused as pre-trained representations for novel text-based tasks. As we have seen, word embeddings have been used to study and use ostensible semantic relationships, i.e., analogies and synonymy clusters, manifested by the usage patterns for practical tasks and applications. I have then shown before that we must carefully consider the representation of the ostensible semantic relationships, given the limits the system exhibits under the test of the private language argument.

The opacity problem[36] in LLMs, as articulated by Bengio et al. (2013), involves the challenge of disentangling the hidden layers of neural networks to understand how they represent and manipulate data. This issue is particularly pertinent in advanced models like Transformers, which embed text within vast networks of interconnected representations, making it even harder to trace how specific outputs are generated. In this sense, the internal processes of LLMs are analogous to Wittgenstein's private objects: they are inaccessible and, therefore, difficult to ascribe meaning to.

35 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Association for Computational Linguistic, 2019, pp. 1-16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, *Attention Is All You Need*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. pp. 1-15, https://doi.org/10.48550/arXiv.1706.03762.

36 Interestingly, the creation of black boxes in algorithms is not always accidental but sometimes deliberate, as seen in proprietary systems like those of Google or Netflix. These systems intentionally obscure their processes to protect intellectual property, creating a situation where the "beetle" within the box is not entirely private but is still inaccessible to the user. This controlled opacity introduces another layer of complexity, where the inaccessibility is enforced rather than inherent, drawing a parallel to Wittgenstein's idea of language games being public and yet containing private elements.

Wittgenstein's philosophy suggests that language is inherently a shared, public activity, and any attempt to model private experiences (or private meanings) will inevitably fall short. This perspective aligns with the challenges faced by AI developers in creating systems with contextual intelligence, as noted by Hollister, Gonzalez, and Hollister (2017). To achieve human-like understanding, AI systems must incorporate contextual information, which precedes and shapes language use. However, when dealing with private, internal experiences – as in the case of the beetle or the black box – these systems encounter fundamental limitations.

The problem of PLA cannot be solved, and the limits of the language games are demarcated. According to Rudin and Radin (2019, 2): «In machine learning, these black box models are created directly from data by an algorithm, meaning that humans, even those who design them, cannot understand how variables are being combined to make predictions». Take all the Xs as processes within a black box, which seems to behold to the "privacy" of the machine – we cannot understand what is going on there. The analogy between Wittgenstein's scenarios and the black box property of algorithms does not want to play the role of anthropomorphizing the machines, «it makes no sense to ascribe thought or thoughtlessness, understanding, misunderstanding or failure of understanding to machines»[37], but to highlight similar features, similar boundaries in both cases. As Rudin and Radin say, the black box hides the algorithmic processes, so how the variables are combined cannot be understood to make the final prediction. Black boxes limit our ability to understand data processing. The black box contains information we cannot access, as in the beetle case. The VSA proposal we have seen does not overcome this limitation. Lowney, Levy, Meroney, and Gayler (2020) offer indeed an intriguing approach by proposing a method to shadow the meaning of a private object, effectively allowing for multiple interpretations of the "beetle" without a fixed, universal meaning. However, this approach, while innovative, does not fully resolve the Wittgensteinian problem. It acknowledges

37 Peter M. S. Hacker, *Human Nature: The Categorial Framework*, Oxford, Wiley-Blackwell, 2011, p. 34.

that the meaning of a term like "beetle" can vary between individuals, reflecting the private experiences that Wittgenstein argues are incommunicable. Yet, by attempting to model these variations, the approach highlights the limitations of trying to capture the essence of meaning within a computational framework.

The black box problem in LLMs can be seen, in conclusion, as a modern incarnation of the philosophical challenges explored by Wittgenstein and Galileo. Just as Wittgenstein's PLA highlights the inaccessibility of private experiences within a public language, the black box problem underscores the difficulty of accessing and understanding the internal workings of AI models. Galileo's approach to mathematical abstraction, while powerful, similarly reflects the limits of understanding that arise when dealing with complex, opaque systems. Together, these philosophical insights provide a framework for grappling with the challenges posed by contemporary AI and its implications for our understanding of language, meaning, and knowledge.

## 6. Conclusion

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

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

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

**Riassunto** Questo articolo analizza i presupposti filosofici dei sistemi contemporanei di elaborazione del linguaggio naturale e dei large language models (LLMs), concentrandosi sul ruolo della matematica nelle attribuzioni di comprensione alle macchine. Si sostiene che tali attribuzioni presuppongano implicitamente una concezione galileiana della matematica come linguaggio capace di rivelare la struttura oggettiva della realtà. Questa visione viene messa a confronto con la filosofia del secondo Wittgenstein, per il quale matematica e linguaggio sono pratiche umane fondate sull'uso, sul contesto e su criteri pubblici condivisi. L'articolo mostra la rilevanza diretta di questa divergenza per la ricerca sull'intelligenza artificiale. Al centro dell'analisi vi è l'argomento del linguaggio privato (PLA), interpretato come caso limite per la formazione del significato e per la possibilità di modellare computazionalmente la comprensione. Vengono esaminati approcci connessionisti dell'elaborazione del linguaggio naturale (NLP), incluse le tecniche di word embedding e le Vector Symbolic Architectures, come tentativi di codificare il significato in strutture matematiche. Particolare attenzione è dedicata ai modelli che pretendono di simulare fenomeni di linguaggio privato, come il caso del "coleottero nella scatola". Si conclude che tali sistemi, pur performanti, manipolano regolarità linguistiche senza catturare le dimensioni normative e contestuali che caratterizzano la comprensione umana nell'IA contemporanea e nello sviluppo dei sistemi intelligenti avanzati.

**Abstract** This paper examines the philosophical assumptions underlying contemporary natural language processing systems and large language models by focusing on the role of mathematics in claims about machine understanding. It argues that attributions of understanding to LLMs tacitly rely on a Galilean conception of mathematics as a realist language revealing the objective structure of reality. This framework is contrasted with the later Wittgenstein's view of mathematics and language as human practices grounded in use, context, and publicly shared criteria. The paper shows how this divergence is directly relevant to artificial intelligence research. Central to the analysis is Wittgenstein's Private Language Argument, interpreted as a limit case for meaning formation and for the possibility of modeling understanding computationally. Connectionist approaches in natural language processing (NLP), including word embedding techniques and Vector Symbolic Architectures, are examined as attempts to encode meaning through mathematical structures. Particular attention is given to proposals claiming to model private language phenomena, such as the "beetle in the box" scenario. The paper argues that, despite impressive performance, these models manipulate linguistic regularities without capturing the normative and contextual dimensions constitutive of human understanding, highlighting a fundamental gap between AI language systems and human cognition in contemporary machine learning research and AI development.