

Meaning and Agency in Large Language Models

MSc Thesis (*Afstudeerscriptie*)

written by

Mayra Huespe

under the supervision of **Prof. Dr. Arianna Betti** and **Dr. Eduardo Giovannini**, and
submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense: **Members of the Thesis Committee:**

August 28, 2025

Dr. Aybüke Özgün (chair)

Prof. Dr. Arianna Betti

Dr. Eduardo Giovannini

Dr. Tom Schoonen

Dr. Hein van den Berg



INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

Since the emergence of large language models (LLMs), there has been growing interest in the question of whether they produce meaningful outcomes (henceforth, the Problem of Meaning in LLMs). In the literature, “producing meaningful outputs” has been understood either as a manifestation of genuine understanding or successful use of language. Existing approaches often adapt philosophical theories of meaning, originally developed for human speakers, to define criteria under which LLMs can be said to produce meaning. However, this strategy creates a tension: we can either support the intuition that LLMs engage in basic linguistic interactions, at the cost of assuming that they simulate human processes of meaning production, or we can deny this intuition and thereby lose the conceptual space needed to account for their distinctive forms of linguistic engagement.

I resolve this tension by proposing a framework that explains how LLMs participate in linguistic interactions without presupposing human-like processes of meaning production. To this end, I first situate the Problem of Meaning within the broader debate on AI agency, adopting Floridi’s distinction between cognitive and non-cognitive views, and show that existing accounts presuppose either a cognitive or soft-cognitive notion of agency. This opens space for a non-cognitive approach. I then develop the first such account of the Problem of Meaning in LLMs, shifting the focus from whether LLMs produce meaningful outcomes to conceptualizing their distinct non-cognitive modes of linguistic engagement within the environment. Philosophical theories of meaning, thus, serve not to define criteria for successful meaning production in LLMs, but rather to conceptualize the linguistic environments in which they operate. In this way, although I still address the Problem of Meaning in LLMs by adopting a philosophical theory of meaning, in particular Putnam’s semantic externalism, I now guide its application through the adoption of a non-cognitive conception of AI agency. Hence, instead of asking whether these models engage in linguistic interactions by simulating a cognitive process, the non-cognitive perspective, along with Putnam’s theory of meaning, shifts the challenge to describing their specific modes of engagement within the linguistic community.

Key Words: AI Agency, LLMs, Non-cognitive Agency, Semantic Externalism, Language Technologies.

Acknowledgements

First and foremost, I would like to thank my supervisor, Arianna Betti, whose guidance and supervision made this thesis possible. I am especially grateful for your methodological advice and insightful feedback, which significantly improved the conceptual clarity of my initial intuitions on this topic. I am equally thankful to my co-supervisor, Eduardo Giovannini, for our many discussions on the topic and for your continuous encouragement over the past months. I also wish to thank the members of my thesis committee, Tom Schoonen and Hein van den Berg, for their constructive suggestions, and Aybüke Özgün for chairing the defense. I am also grateful to Adriana Gonzalo and Ezequiel Zerbudis for their thoughtful feedback in various sections of this thesis.

My sincere thanks also go to Giorgio Sbardolini, who kindly supervised an individual research project on LLMs and linguistic representations. This project introduced me to the preliminary questions that later developed into the core intuitions motivating this thesis. Throughout the MoL, I found both research projects and other more informal discussion spaces to be highly stimulating environments for the exchange of ideas, in which I had the chance to learn not only about my own interest, but also about the exciting works that my peers were pursuing. I am grateful to the ILLC and the MoL community for continually fostering such stimulating environments. My sincere thanks go as well to Dingmar van Eck, Tanja Kassenaar and Ines Crespo, for their valuable advice that made my MoL journey much smoother and more enjoyable.

On a more personal note, this thesis would not have been possible without the love, support, and encouragement of my family and friends. To my friends, especially my lifelong ones in Argentina, Lula, Lauri, Colo, Nati, Milo, Chuchi, Beyabel, Josu, Azu, and Ninu, although I couldn't be there for many important moments in recent years, you always found ways to make me feel close and included. I will always be grateful for that and for your invaluable friendship. To the friends I had the pleasure of meeting during the MoL, Michael, Dorielle, Alyssa, and Sara, thank you for making my time in Amsterdam so memorable, with our ramen tours, walks around the city, and study sessions that made everything feel lighter. To my family, Vilma, Alfredo, and Iván: thank you for always supporting my decisions and encouraging me in all my dreams.

Last but not least, I want to thank my partner, Arunavo, for our conversations, in the Pink Room, at the end of the well-known second-floor corridor, and on our many bike trips. Thank you for showing me that philosophy can be discussed even while climbing a bridge in the rain, and for every culinary adventure we shared together. Because of you, Amsterdam will always remain a *pink* memory.

Contents

Introduction	2
2 The Problem of Meaning in LLMs	13
2.1 Preliminaries	13
2.2 A Non-cognitive Notion of Artificial Agency	16
2.2.1 Cognitive and non-cognitive perspectives of AI	17
2.2.2 The cleaving power of AI	19
2.2.3 Decoupling agency from intelligence	19
2.3 The Problem of Meaning in LLMs	21
2.3.1 The roots of the problem of meaning in LLMs	22
2.3.2 Addressing the Problem of Meaning in LLMs	24
2.3.3 A technical introduction to LLMs:	27
2.4 Cognitive and Non-cognitive Approaches to the Problem of Meaning	32
2.4.1 Mentalist and non-mentalist theories of meaning	32
2.4.2 A Defense of non-cognitive approaches	35
3 A Non-cognitive Approach to the Problem of Meaning in LLMs	41
3.1 Preliminaries	41
3.2 Semantic and Metasemantic Methodologies	44
3.2.1 Semantic and metasemantic theories of meaning	44
3.2.2 Semantic and metasemantic methodologies	45
3.2.3 Two Semantic attempts to address the Problem of Meaning in LLMs	46
3.3 Ambitious and Modest Scope of Application of Putnam’s Theory of Meaning	52
3.3.1 Twin Earth thought experiment	52
3.3.2 Two interpretations of the Twin Earth thought experiment	53
3.3.3 Why Language does not need to hook onto the world	54
3.4 A Metasemantic and Modest Application of Putnam’s Theory of Meaning	56
3.4.1 Extensions: the meaning of “meaning” in the age of LLMs	57
3.4.2 Individual Capabilities: the role of LLMs in the linguistic community	60
Conclusion and Future Work	65
Bibliography	69

Introduction

Since the emergence of Large Language Model-based technologies capable of creating new text, such as ChatGPT, in the philosophy of AI literature, there has been growing interest in the extent to which these technologies truly understand or use language. Large language models (LLMs) are deep neural networks trained on massive amounts of text to predict the next word given an initial sentence (Zhao et al., 2024). These models generate word by word or, more precisely, token by token, new, coherent text based on an initial input sequence. One of the concerns that has been taking a prominent role in the literature is what I propose to call the Problem of Meaning in LLMs (for short, The Problem), which can be formulated as follows:

The Problem: Do LLMs produce meaningful outcomes?

In the literature, producing meaningful outcomes has been understood as a manifestation of either genuine understanding or the successful use of language. A prominent approach that has gained attention over the past years addresses this question in three steps: (i) adopting philosophical theories of meaning developed initially to explain meaning production in human speakers, (ii) defining criteria for when meaningful expressions can be said to be successfully produced, and (iii) applying these criteria to LLMs to determine whether they meet that standard (Bender et al., 2021; Grindrod, 2024; Gubelmann, 2024a; Mandelkern & Linzen, 2024).

This strategy frames The Problem as a tension between two intuitions: The first one relates to the interaction between humans and language technologies. From offering movie recommendations or suggesting vegetarian recipes to providing medical advice or recommending mental health treatments, LLMs are being increasingly embedded into technologies we use in our everyday lives (Biswas, 2023; Bloch & Shaul, 2024; Cascella et al., 2023; Haltaufderheide & Ranisch, 2024; Hua et al., 2025; Yang et al., 2023). We interact with LLM-based technologies, understanding the text they generate, and acting accordingly (see Table 1.1). The first intuition, hence, is that we successfully engage in linguistic interaction with such technologies. I propose to call this intuition *LLMs as artificial speakers*.

However, and this is where the second intuition comes into play, these technologies remain far from resembling the cognitive capabilities underlying human production of meaning. Interacting with ChatGPT makes this clear: it frequently makes factual errors, struggles with basic mathematical reasoning, and generates text that appears plausible but contains incorrect information, a phenomenon known as hallucinations (see Table 1.2) (Arkoudas, 2023; Borji, 2023; Christian, 2023; Cobbe et al., 2021; Hua et al., 2025; Perez et al., 2023; Rumbelow, 2023; Y. Zhang et al., 2025). These frequent and well-known failures highlight their genuine lack of linguistic understanding, semantic knowledge, capacity to refer, consciousness, intentionality, among other things that

characterize human production of meaning (see Table 1.3) (Bishop, 2021; Floridi, 2023a, 2023b, 2025).



Speaker	Interaction
 Patient (Jennifer, 39F)	I have had a cough for two weeks. What could it be?
 AI Agent	<p>Hi Jennifer, a cough that lasts for two weeks could be due to various factors. Based on your medical history and the information provided, here are some possible causes to consider:</p> <ul style="list-style-type: none"> • Asthma: Given your history of asthma, it’s possible that your cough is related to poorly controlled asthma or an asthma exacerbation. Asthma-related coughs can worsen with exposure to cold air or irritants. • Postnasal drip: When your nose or sinuses produce mucus, it can drip down the back of your throat and trigger a cough reflex. This is known as postnasal drip or upper airway cough syndrome (UACS). • Acid reflux (GERD): Gastroesophageal reflux occurs when stomach acid flows back into the tube connecting your stomach and throat. This can lead to chronic coughing. • Chronic bronchitis: Defined as a productive cough that lasts at least three months, with recurring bouts for at least two consecutive years. Often caused by cigarette smoking or environmental irritants. • Other possibilities: Chronic cough can also result from allergies, lung infections, or certain medications. <p>Recommendation: Based on the information provided, it would be advisable to schedule an appointment with your doctor to discuss your symptoms and undergo a thorough evaluation. They can provide a more accurate diagnosis and recommend appropriate treatment options based on your specific situation.</p>

Table 1.1: Example that supports the intuition that LLM-based technologies, such as in this case, a chatbot released by K Health, engage in successful linguistic interactions with the user. This interaction constitutes a sample of how the new K Agent answers questions based on the patient’s medical history and electronic medical record. (Bloch & Shaul, 2024).

Although none of these limitations undermines the remarkable advances they introduced in the field of language technology, nor do they prevent their integration of such tools into the automation of increasingly complex tasks (see, for instance, the recent launch of OpenAI, 2025), such achievements seems to occur despite the absence of a cognitive process of meaning production. The second intuition, thus, is that these technologies do not simulate or resemble the cognitive capabilities underpinning human production of meaning. I refer to this second intuition as *LLMs as no smarter than a toaster*, based on the somewhat unexpected but recurring analogy in the literature of these models’ intelligence with that of a toaster (Barry, 2025; Floridi, 2023a, 2023b).

These two intuitions seem to trace a forking conceptual path. On the one hand, adopting a positive attitude towards the intuition that these technologies lack the cognitive capabilities underpinning human production of meaning seems to require setting aside the idea that these

technologies are artificial speakers. On the other side, exploring how such technologies engage in linguistic interaction seems to require starting from a rejection of the idea that LLMs are no smarter than a toaster. Put differently, the price for adopting an optimistic view of LLMs-based technologies as engaging in successful linguistic interactions with human speakers seems to be adopting a similarly optimistic stance regarding their capacity to mimic cognitive capabilities in the production of meaning. Similarly, adopting a skeptical view of this latter capacity undermines the foundation necessary for conceptualizing how these technologies successfully engage in linguistic interactions. Thus, it seems we either prioritize the intuition that LLM technologies are artificial speakers, at the cost of presupposing strong assumptions about these models’ capacity to resemble cognitive capabilities in the production of meaning, or we provide a positive account of why LLMs are technologies no smarter than a toaster, at the expense of understanding how these technologies actually engage in linguistic interactions.



Speaker	Interaction
 User	Is 9677 a prime number?
 GPT-4	No, 9677 is not a prime number. It can be factored into 13 and 745, as $9677 = 13 \times 745$.

Table 1.2: Illustrative example supporting the intuition that LLM-based technologies, such as GPT-4, lack genuine linguistic understanding. This case represents the phenomenon known as *snowballed hallucination*. GPT-4 mistakenly asserts that 9677 is not a prime number, and it further incorrectly explains that $13 \times 745 = 9677$. Snowballed hallucinations, however, do not indicate that the model lacks the relevant piece of information, since when asked separately, the model correctly identifies 9677 as a prime number. Instead, these failures reflect a tendency to preserve internal consistency with an earlier hallucination (M. Zhang et al., 2024).

This leads to the following dilemma: Should we focus on demonstrating how far these technologies are from simulating a cognitive process of meaning production and, in consequence, how far they are from successfully participating in linguistic interactions? Or should we instead devote our conceptual efforts to understanding LLMs-human linguistic interactions as they occur, even if this involves making strong assumptions about LLMs’ cognitive capabilities? In the recent philosophical and computational linguistic literature on The Problem, the positions adopted often shift back and forth between these seemingly opposing intuitions (see, for instance, Bender & Koller, 2020; Bender et al., 2021; Grindrod, 2024; Gubelmann, 2024b; Mandelkern & Linzen, 2024; Piantadosi et al., 2024).

On the one hand, some authors provide conceptualizations of how these technologies successfully participate in linguistic interaction, but often at the cost of assuming strong presuppositions about the models’ capacity to mimic human capacity for meaning production (Lederman & Mahowald, 2024; Mollo & Millière, 2023; Piantadosi et al., 2024). On the other hand, others chose to emphasize how far these language technologies are from simulating those cognitive capabilities typically associated with the production of meaning, rejection, in consequence, the intuition that LLMs participate in linguistic interactions (Bender & Koller, 2020; Bender et al., 2021;

Bottazzi Grifoni & Ferrario, 2025; Gubelmann, 2024a).








Speaker	Interaction
 GOOGLE AI	What can I help with? A brief explanation is all I need.
 User (Dave Barry)	I'm Dave Barry. The Google AI Overview for me says that I'm dead, and that when I was alive I was a political activist in Dorchester. Neither of these things is true. I am not dead, and I am not a political activist in Dorchester.
 GOOGLE AI	Sorry, I didn't understand your question. Can you try rephrasing it? Short phrases work best.
 User (Dave Barry)	I'm Dave Barry. I am alive. Google AI says I am dead. I am not dead. Google AI also says I was a political activist in Dorchester. I was not. I still am not.
 GOOGLE AI	Apologies, I'm still having trouble understanding you. Let's give it one more try. Please describe your issue in the text box below.
 User (Dave Barry)	My issue is that Google AI says I am dead, but I am not.
 GOOGLE AI	Alright, here are a few options I found: <ul style="list-style-type: none"> • Resolve Reviewed Claims • AI can and will make mistakes • Why connections between the search and the result may seem unclear • How search results work • Something else

Table 1.3: Another example that illustrates the second intuition comes from Dave Barry. He discovered that the Google AI Overview incorrectly stated he was dead, and he attempted to correct this error. After this frustrating back-and-forth, Barry remarked: *It was like trying to communicate with a toaster. Actually, it was worse than that. I bet if I stood in front of a toaster and shouted "I'm not dead" at it enough times, eventually it would respond "Oh! I get it now! You're not dead!" Whereas Google AI seemed incapable of grasping this point.* (Barry, 2025) .

Rather than adopting one of the horns of this dilemma, this thesis develops a conceptual framework that reconciles these two intuitions. In particular, this thesis aims to articulate a conceptual framework that enables the exploration and conceptualization of how LLM technologies successfully engage in linguistic interactions, while maintaining a skeptical view regarding whether they can replicate the cognitive process behind the production of meaning. This goal thus can be stated as follows:

Goal: To elaborate a novel framework for examining how LLMs can participate in linguistic interaction free from cognitive presuppositions.

Schematically, to address this apparent dilemma, the strategy I propose involves two fundamental steps: first, I situate The Problem as a specific topic within the broader debate on AI agency (*Specific Goal 1*); second, I show how this framing provides an ideal conceptual space for

reconciling both intuitions (*Specific Goal 2*).

More specifically, I begin by connecting two debates that, to the best of my knowledge, have been developed independently in the philosophical literature: the discussion on AI agency, framed in terms of cognitive and non-cognitive perspectives, on the one hand, and The Problem, on the other.

To connect these two debates, I adopt Floridi's reconstruction of the discussion about agency (Floridi, 2023b, 2025). In this account, two fundamental concepts of agency are distinguished: a cognitive and a non-cognitive notion. On the one hand, from a cognitive perspective, investigating AI agency requires analyzing *whether* these technologies replicate human intelligence. On the other hand, from a non-cognitive perspective, studying AI agency involves analyzing *how* AI technologies replicate human behavior through non-cognitive means. In other words, this second notion explores how AI technologies reproduce human-specific modes of engagement with the environment by non-cognitive means. This shift involves moving from the attempt to determine whether there is something (i.e., cognitive-like features) in AI technologies that justifies the claim that they have agency, to the recognition that these technologies already exhibit non-cognitive agency.

Floridi (2025) provides two arguments in favor of the non-cognitive notion of agency. First, he appeals to parsimony: a non-cognitive account offers a framework for investigating agency with a lower theoretical burden, as it does not require attributing to AI technologies strong assumptions about their ability to simulate human cognition. Second, this shift enables a focus on the transformative impact of these technologies. In particular, it offers a perspective from which to examine how such technologies modify the environment and the tasks performed within it by introducing new agents that, through non-cognitive means, carry out tasks previously accomplished by cognitive agents.

I then turn to The Problem. I show that the literature on this topic can be broadly divided into mentalist (e.g., Bender et al., 2021; Piantadosi et al., 2024) and non-mentalist approaches (e.g., Grindrod, 2024; Gubelmann, 2024b), depending on the extent to which the underlying theory of meaning relies on mental or psychological constructs of speakers to explain meaning. Based on this distinction, I make explicit the connection between the debate on agency and the debate on The Problem. In short, I argue that mentalist positions on meaning presuppose a cognitive conception of AI agency, while the non-mentalist position presupposes a *soft* cognitive notion. On this basis, I show that so far, no approach to the Problem has adequately been grounded in a non-cognitive notion of AI agency.

More specifically, I show that mentalist theories of meaning expect LLMs to produce meaning by simulating or resembling the cognitive processes underlying human meaning production. Non-mentalist theories, by contrast, take an initial step toward grounding their approach in a non-cognitive notion of AI agency by rejecting the framing of The Problem in terms of whether LLMs replicate cognitive processes of meaning production. Instead, they ask whether LLMs replicate human modes of linguistic engagement with the environment. However, I argue that these approaches still fall short of fully committing to a non-cognitive notion of AI agency. Such a commitment would require shifting the question from whether to how LLMs are actually replicating human modes of linguistic engagement through non-cognitive means. I refer to this

intermediate position as one grounded in a soft cognitive notion of AI agency.

Through this argument, I make explicit that, so far, there is no non-mentalist approach to The Problem that is properly grounded in a non-cognitive notion of AI agency. Thus, the first Specific Goal is:

Specific Goal 1 (corresponding to Chapter 1): To show that mentalist approaches to The Problem presuppose a cognitive notion of AI agency, while non-mentalist ones presuppose a soft cognitive notion, thereby making explicit the lack of an approach grounded in a non-cognitive notion.

Based on this reconstruction, in the second chapter, I address the gap identified in Chapter 1 by developing an approach to The Problem grounded in a non-cognitive notion of AI agency. I begin by showing how the commitment to a soft cognitive notion of AI agency is manifested in a specific non-mentalist case: Mandelkern and Linzen (2024)'s proposal to address The Problem by adopting Putnam's externalism. For the sake of brevity, Putnam's externalism can be summarized by the slogan: the linguistic community fixes the meaning of our words. Based on this theory of meaning, Mandelkern and Linzen (2024) frame The Problem in terms of whether LLMs successfully engage with the linguistic community by replicating human modes of linguistic interaction. On the one hand, by adopting a non-mentalist theory of meaning, they take a first step toward a non-cognitive notion of AI agency. In particular, they do so by no longer asking whether LLMs replicate a cognitive process of meaning production. However, since they still frame The Problem in terms of whether LLMs replicate human modes of linguistic engagement with the environment (in this case, the linguistic community), they remain committed to what I call a soft cognitive notion of AI agency. In the first part of the second chapter, I start by reconstructing the interpretive decisions that led Mandelkern and Linzen (2024) to frame The Problem as grounded in a soft cognitive notion of AI agency.

In the second part of the second chapter, in contrast to Mandelkern and Linzen (2024)'s proposal, I suggest a novel application of Putnam's externalism to address The Problem, guided by a non-cognitive notion of agency. I begin by making explicit my departure from the interpretive decisions that led Mandelkern and Linzen (2024) to frame the problem in cognitive terms. I then reframe The Problem in terms of *how* LLMs are linguistically engaging with the linguistic community. In particular, I propose to examine their integration into the linguistic division of labor within the linguistic community. Hence, rather than analyzing whether there is something in these models that justifies their belonging to the linguistic community, I propose asking how the linguistic community is actually integrating them. Accordingly, the second *Specific Goal* can be stated as follows:

Specific Goal 2 (corresponding to Chapter 2): Conceptualizing The Problem of Meaning in LLMs in terms of an exploration of how these technologies are being integrated into our linguistic communities.

As an important remark on the scope of this thesis, I do not aim here to solve this version of The Problem, but rather to suggest a potential new way of understanding it. The novelty

of this proposal consists in elaborating a conceptual framework for The Problem that allows us to maintain both skepticism about the cognitive capabilities of LLM-based technologies and an optimistic view of their ability to engage in linguistic interactions. Fundamentally, this conceptual space emerges, first, by contextualizing The Problem as a specific discussion within the broader topic of AI agency, and then by adopting both a non-cognitive notion of AI agency and a non-mentalist theory of meaning: in this specific case, Putnam’s Externalism. Thus, LLMs can be artificial speakers, despite being technologies no smarter than a toaster. The keystone for reconciling these two intuitions is to stop asking whether LLMs can simulate the cognitive production of meaning, or even whether they can replicate human modes of linguistic engagement with the environment. Instead, the focus should shift to describing the non-cognitive modes of linguistic engagement through which these technologies are being integrated into linguistic communities.

In the sections that follow, I outline the steps undertaken to prove goals (1) and (2). Each goal is addressed in a separate chapter, forming two independent contributions that together support the broader research objective of this thesis, namely, elaborating a framework in which it is possible to reconcile the two relevant intuitions.

Part 1: Agency in LLMs

In what follows, I enumerate and briefly describe the steps I follow to achieve the first specific goal of this thesis: To show that mentalist approaches to The Problem presuppose a cognitive notion of AI agency, while non-mentalist ones presuppose a *soft* cognitive notion, thereby making explicit the lack of an approach grounded in a non-cognitive notion.

***Step 1.1:** Describing the two perspectives that frame the debate about agency, namely, the cognitive and non-cognitive views*

I begin by adopting Floridi’s framework for agency, in which he develops a non-cognitive notion of agency in response to a cognitive one. According to Floridi, AI has been developed along two main lines of research: cognitive and non-cognitive. The first treats AI as a branch of cognitive science, concerned with developing technologies that replicate human intelligence through non-biological means. The second views AI as a branch of engineering, focused on simulating human behavior. These two conceptions endorse different notions of agency. From the cognitive perspective, the extent to which AI technologies instantiate agency depends on how well they replicate human intelligence. From the non-cognitive perspective, AI agency is free from cognitive constraints, as it is understood in terms of how these technologies are simulating, by non-cognitive means, the specific ways in which humans interact with their environment while doing a particular task.

***Step 1.2:** Mapping the literature on The Problem and making explicit the common strategy that underlies the various proposed solutions.*

In the second part of this chapter, I focus on The Problem. This problem is typically framed in the literature through the following general question: Do LLMs generate meaningful outcomes? In

this chapter, I offer, to the best of my knowledge, the first comprehensive and systematic map of the various attempts to answer this question in both the philosophy of AI and the computational linguistics literature. This systematization makes explicit the common strategy underlying the different solutions proposed in both fields to address The Problem in LLMs. This strategy, which I call the *three-step* framework to address The Problem, consists of the following stages: (i) adopting philosophical theories of meaning that were developed initially to explain meaning production in human speakers, (ii) defining criteria under which meaning is successfully produced; and (iii) applying these criteria to assess whether the outputs of LLMs meet that standard.

Step 1.3: *Characterizing the literature on The Problem of Meaning in LLMs by distinguishing between mentalist and non-mentalist positions.*

Based on the three-step framework, I develop a novel characterization of the different approaches to The Problem. This characterization hinges on whether the approach undertakes step (1) by adopting either a mentalist or a non-mentalist theory of meaning. On the one hand, mentalist approaches adopt theories in which meaning is explained—partially or entirely—in terms of the mental or psychological features of speakers. Non-mentalist approaches, on the other hand, avoid attributing explanatory power to these aspects of the speaker. Instead, they explain meaning in terms of pragmatic norms, histories of use, naming practices, and other constructs that emphasize the linguistic environment to which speakers belong over their mental or psychological features.

Step 1.4: *Showing that mentalist approaches to The Problem presuppose a cognitive notion of agency, and non-mentalist approaches presuppose a soft cognitive notion of agency, thereby making explicit the lack of an approach grounded in a non-cognitive notion.*

In this section, I demonstrate that mentalist approaches ground The Problem as a specific instance of the broader investigation into cognitive agency, while non-mentalist perspectives take a step closer to treating it in terms of a non-cognitive notion of AI agency. However, these approaches still cannot be said to be fully grounded in a non-cognitive notion of AI agency à la Floridi. In doing so, I articulate the first key contribution of this chapter: establishing a connection between the debate on AI agency and the discussion about The Problem, and making explicit the gap that I address in Chapter 2.

Step 1.5: *Arguing in favor of non-mentalist approaches*

I finish the first chapter by providing an argument in favor of non-mentalist approaches to The Problem. In particular, I show that of the two advantages Floridi ascribes to a non-cognitive framework for agency, non-mentalist approaches grounded in a soft notion of AI agency already enjoy one of them. Specifically, I argue that these approaches benefit from leaving behind the theoretical burden of studying agency in cognitive terms. However, they still do not enjoy the second advantage: namely, providing a more effective perspective from which to examine how such technologies modify the environment and the tasks performed within it by introducing new agents that, through non-cognitive means, carry out tasks previously accomplished by cognitive

agents.

Part 2: Meaning in LLMs

In the second chapter, I adopt a non-cognitive notion of agency along with Putnam’s externalism as a non-mentalist theory of meaning, and propose conceptualizing The Problem based on both. I begin by analyzing a previous attempt to solve The Problem developed by Mandelkern and Linzen (2024), and I show why it is grounded in a *soft* cognitive notion of AI agency. I reconstruct two interpretive decisions that lead them to rely on this cognitive notion: the adoption of a semantic rather than a metasemantic methodology, and an ambitious rather than a modest application of Putnam’s theory of meaning. Grounded in my explicit commitment to a non-cognitive notion of agency, I depart from these two features by adopting a metasemantic and modest application. This leads me to offer the first attempt to frame The Problem by applying Putnam’s theory of meaning in a way that is grounded in a non-cognitive notion of AI agency. In what follows, I enumerate and briefly describe the steps I follow to develop this novel approach to The Problem.

Step 2.1: Introducing the distinction between semantic and metasemantic methodologies for approaching The Problem, and showing that Mandelkern and Linzen, 2024 adopts the former.

To clarify this distinction, I begin by introducing the difference between semantic and metasemantic theories of meaning. In short, semantic theories focus on ascribing meaning to expressions, while metasemantic theories focus on the facts in virtue of which we ascribe meaning to those expressions. A semantic methodology, then, addresses The Problem by asking what the meaning of LLMs’ outputs is. I show how this methodology was first implemented by one of the foundational works on the topic, Bender et al. (2021), which laid the groundwork for subsequent discussions, including the work of Mandelkern and Linzen (2024). A metasemantic methodology, by contrast, aims to explain the facts in virtue of which LLMs’ outcomes are meaningful.

Step 2.2: Introducing the distinction between ambitious and modest applications of Putnam’s Externalism to address The Problem, and showing that Mandelkern and Linzen, 2024 follows the former.

Mandelkern and Linzen (2024) applies Putnam’s Externalism as a theory that can determine whether, and how, LLMs’ outcomes might be hooked into the world. This application is inherited from an interpretation of this theory as addressing the well-known challenge in philosophy of language of *how words hook into the world*. The intuition behind this application is that if Putnam’s theory of meaning was intended to explain how language hooks into the world, it can also be used to determine whether LLMs’ outcomes hook into the world. In the second Chapter, I show how this interpretation aligns more closely with Putnam’s earlier writings on meaning. I suggest calling the application that follows from this interpretation an ambitious application of Putnam’s externalism.

In contrast to this ambitious application, I propose a more modest one, grounded in Putnam’s later rejection of the challenge of how language hooks into the world. Following Putnam in later writings on meaning, this challenge only arises when we overlook the role of labor distribution in

fixing meaning within a linguistic community. Once we adopt a view of meaning that accounts for how different groups of experts fix meanings, the challenge no longer holds as a valid one. Thus, I show how a modest application of this theory of meaning involves shifting the focus from explaining how LLMs' outputs hook into the world to describing their actual modes of linguistic engagement with the linguistic community.

Step 2.3: *Showing why a semantic and ambitious application of Putnam's Externalism contributes to framing The Problem in a way it presupposes a soft cognitive notion of agency.*

The route that the Mandelkern and Linzen, 2024 follow can be summarized as follows: given their semantic methodology, they start from the general question of whether LLMs' outcomes mean something. Then, because of their ambitious application, they conceptually refine The Problem as the question of how these outcomes hook into the world. Accordingly, the authors claim that LLMs' outcomes would mean something if these outcomes were hooked into the world in a similar manner as speakers' words are hooked into the world. Speakers' words are connected to the world by virtue of their membership in the linguistic community, more specifically, given their capacity to generate meaningful words in accordance with the history of usage present in that community. Hence, the authors frame The Problem in terms of whether these models can simulate the internal features that underlie a speaker's capacity to generate a word in accordance with its history of use in the linguistic community. In other words, a semantic along with an ambitious application of Putnam's theory of meaning contributes to frame the problem in terms of the following question: *Do LLMs successfully engage with the environment (in this case, the linguistic community) by replicating human modes of linguistic engagement (in this case the speaker's capacity to generate a word in accordance with its history of use in the linguistic community)?*. This question, as remarked, presupposes a *soft* cognitive notion of agency.

Step 2.4: *Elaborating a non-cognitive, that is, a metasemantic and moderate, application of Putnam Externalism to account for how LLM-based technologies successfully engage in linguistic interactions.*

Based on adopting a metasemantic method and a modest application of Putnam's externalism, in the final section of Chapter 2, I frame The Problem as follows: *How are LLMs linguistically engaging with the environment (in this case, the linguistic community)?* As remarked, this way of framing the problem involves grounding my proposal in a non-cognitive notion of AI agency *à la Floridi*. I finish, thus, by providing a first attempt to answer this question based on the two dimensions of meaning that are relevant within Putnam's theory of meaning: extension and individual capabilities.

The first dimension concerns the extension that the community ascribes to a word, which the relevant group of experts within that community fixes. By applying this dimension to LLMs, I suggest exploring how various expert groups define the technical terms *meaning* in LLMs within NLP and computational linguistics research. I propose that Explainable AI (xAI) techniques may play a fundamental role in fixing the meaning of *meaning* in LLMs. This leads one to ask how LLMs are being integrated into the linguistic community, particularly in terms of how these language technologies are introducing a new area of expertise that fixes the meaning of *meaning*

from a new perspective.

The second dimension involves the notion of individual capabilities and the phenomenon of semantic deference. In particular, the slogan that *the meaning is fixed by the linguistic community* also involves a second dimension, namely, individual capabilities. The linguistic community is organized around the distinction between expert and non-expert. Fundamentally, when a speaker assumes the role of an expert in an interaction, non-experts tend to defer to them semantically. This is the so-called phenomenon of semantic difference. Based on this second dimension of meaning, I suggest that another way in which we can explore how LLMs are being integrated into the linguistic community is by studying the role they play within the expert-non-expert dichotomy. This conceptualization leads to asking how the introduction of these new technologies is potentially transforming the linguistic community by introducing a new artificial speaker to whom non-expert speakers might be deferring.

Step 2.5: Making explicit how this approach to The Problem reconciles the two intuitions.

I conclude by showing how this non-mentalist approach to The Problem fully enjoys the two advantages that Floridi ascribes to a non-cognitive notion of AI agency. It not only provides a framework for investigating agency with a lower theoretical burden, but also offers an effective perspective from which to examine the transformative impact of these technologies. In particular, shifting the focus from the question of whether LLMs instantiate linguistic agency to the question of how they are linguistically engaging within the linguistic community allows us to focus on their transformative role. This impact arises from both their potential to reshape what the linguistic community understands by “meaning” and their capacity to alter the network of semantic deference between expert and non-expert speakers.

This framework thus opens a conceptual space in which our two intuitions can be reconciled: LLMs can be seen as artificial speakers insofar as they engage linguistically with the environment in the modes defined above—yet these modes of engagement are achieved through non-cognitive means. Indeed, their transformative potential arises precisely from the fact that they participate in linguistic practices without simulating any cognitive process. In other words, although these technologies may be “no smarter than toasters”, they still interact with (and potentially modify) the linguistic community.

Chapter 2

The Problem of Meaning in LLMs

2.1 Preliminaries

In the present chapter, I aim to connect the debate on AI agency (cognitive vs. non-cognitive views) with the discussion on The Problem (mentalist vs. non-mentalist views). In particular, I show that while mentalist approaches presuppose a cognitive notion of AI agency, non-mentalist approaches take a step closer to a non-cognitive conception. Still, non-mentalist approaches rely on what I propose to call a *soft* cognitive notion of AI agency. Accordingly, I conclude by claiming that there is conceptual space for addressing The Problem by adopting both a non-mentalist theory along with a non-cognitive notion of agency, the objective I pursue in Chapter 2.

I begin this argument by presenting Floridi’s recently developed novel approach to Artificial Agency (AA). This approach is grounded in the rejection of the cognitive perspective on AI (Floridi, 2023b, 2025). Since its early days, two fundamental research lines have defined goals, interpreted the achievements, and guided the development of AI: the cognitive and the non-cognitive research lines. The first one understands AI as a branch of cognitive science focused on replicating *human intelligence* by non-biological means. The second research line, the non-cognitive, understands AI as a branch of engineering focused on developing technologies that replicate *human behavior*. In other words, rather than asking whether AI technologies replicate human intelligence, the latter approach focuses on studying to what extent these technologies replicate human behavior by non-cognitive means.

Moving towards the concept of agency, from a cognitive perspective, AI agency is understood in terms of how effectively these technologies replicate human intelligence. The relevant question is whether, and how, the concept of intelligence can be extended to include under its scope AI technologies. This is the so-called Artificial Realisability of Intelligence Thesis (ARI) (Floridi, 2025). Floridi’s account of artificial agency is grounded in rejecting the cognitive perspective and, consequently, in rejecting framing AI agency through the ARI thesis. Rather than asking *whether* AI technologies instantiate agency by successfully simulating intelligence, this non-cognitive notion involves asking *how* AI technologies are instantiating agency by non-cognitive means. This is known as the Artificial Realisability of Agency (ARA) thesis. Two fundamental reasons underpin this shift. First, studying AI agency in cognitive terms entails a heavier theoretical burden. Second, once we set aside this burden, we gain the conceptual space to focus on how AI technologies, while doing tasks that used to be accomplished by cognitive agents, are transforming

both the nature of the tasks and the environments in which those tasks are performed. In other words, adopting a non-cognitive notion of AI agency offers a more effective framework for examining the impact of AI technologies.

Having established the framework for an agency that underpins this chapter, the next stage of the argument involves introducing The Problem. Recently, LLMs and other generative AI systems have taken on a prominent role in the philosophical literature (Bottazzi Grifoni & Ferrario, 2025; Cappelen & Dever, 2021; Floridi, 2023a; Gubelmann, 2024a; Lederman & Mahowald, 2024). A fundamental question that has gained increasing attention is whether LLMs can produce meaningful outcomes (Bender & Koller, 2020; Grindrod, 2024; Mandelkern & Linzen, 2024). This concern traces back to the Grounding Problem in symbolic AI, where the challenge was understanding how purely symbolic systems could ground, or "hook" their symbols onto the world to acquire meaning (Harnad, 1990). That is to say, the question was how such symbols could be bearers of meaning if they were defined and manipulated only by the relevant logical rules of the systems, lacking any ground in anything beyond the systems themselves. LLMs, however, belong to a different paradigm: statistical AI. Their technical foundations differ significantly from symbolic AI systems. Rather than manipulating symbols through logic rules, LLMs are deep neural networks trained on a large amount of text. Fundamentally, these models are trained to predict the next token in a sequence. Yet, despite these technical differences, the underlying philosophical concerns persist, albeit now in the form of the question: (1) *Do LLMs generate meaningful outcomes?* (Bottazzi Grifoni & Ferrario, 2025; Grindrod, 2024; Gubelmann, 2024a; Lederman & Mahowald, 2024; Pepp, 2025). In particular, the question is whether aspects such as the training process or the architectural features of these models can justify assigning meaning to the output generated through next token prediction.

In recent years, there has been a growing body of literature in both computational linguistics and philosophy of AI dedicated to addressing The Problem. A key step in the argument I develop in this chapter involves distinguishing between approaches to The Problem that rely on mentalist theories of meaning and those that depend on non-mentalist theories. To lay the groundwork for this distinction, I first provide a comprehensive and systematic review of the relevant literature, highlighting the common strategy employed to address The Problem. This survey shows that most approaches follow a three-step structure. The first step consists of adopting a theory of meaning. For example, Piantadosi et al. (2024) adopts conceptual role semantics, as developed by Block (1998), and Gubelmann (2024b) adopts a version of inferentialism, based on the work of Brandom (1994, 2010, 2021). The second step involves deriving, from the adopted theory of meaning, a condition that LLMs must satisfy in order to generate meaningful outcomes. If one follows conceptual role semantics, where the meaning of a symbol is understood, roughly speaking, in terms of the conceptual role of a mental representation, then LLMs would generate meaningful outcomes if they can internally represent conceptual roles (Piantadosi et al., 2024). Alternatively, if meaning is understood in terms of the pragmatic norms that regulate language use in communicative contexts, a positive answer to The Problem would depend on whether LLMs can encode during training those pragmatic norms (Gubelmann, 2024a). The third and final steps consist of investigating whether LLMs satisfies the relevant condition derived from the adopted theory of meaning. This is typically approached by analyzing specific aspects of LLMs'

training or architectural features. For example, Piantadosi et al. (2024) provides a positive answer by showing that vector embedding, a core component of LLMs' architecture, encodes conceptual roles. On the other hand, Gubelmann (2024b) also provides a positive answer, arguing that the vast amount of textual data used to train LLMs is sufficiently rich to capture and encode the pragmatic norms that govern language use. From this perspective, it is plausible to claim that LLMs, through their training, encode such norms (Gubelmann, 2024b).

As this systematic reconstruction of the strategy employed to address The Problem reveals, a fundamental first step involves selecting a theory of meaning. This provides room for characterizing these different solutions based on the adopted theory of meaning. The following relevant step in my argument, therefore, involves distinguishing between attempts grounded in mentalist theories of meaning and those based on non-mentalist theories of meaning. Mentalist theories attribute explanatory power to mental or psychological features of speakers in their accounts of meaning. Non-mentalist theories, on the other hand, explain meaning without appealing to such theoretical constructs. The two proposals mentioned earlier illustrate each side of the distinction. Piantadosi et al. (2024), by adopting the foundational theory of meaning of Block (1998), aligns with the computational theory of mind tradition, in which internal mental representations play a fundamental explanatory role in accounting for meaning. Gubelmann (2024b), by contrast, adopts a theory of meaning rooted in the inferentialist tradition in philosophy of language, which does not rely on mental constructs to explain meaning.

The final phase of the argument I elaborate in this chapter involves connecting the debate about agency with the discussion regarding The Problem. Fundamentally, it involves showing how mentalist approaches presuppose a cognitive notion of AI agency and non-mentalist approaches give the first step towards a non-cognitive notion. However, they still endorse a *soft* cognitive notion of AI agency. More specifically, I begin this final section by showing that under mentalist theories of meaning, The Problem is better understood as an instance of the ARI thesis. Mentalist theories frame meaning production as a fundamentally cognitive process, that is, as a manifestation of human intelligence. As a result, adopting mentalist theories to address The Problem involves understanding LLMs as language technologies whose ultimate goal is to produce meaning by mimicking the cognitive processes of meaning production. The Problem, thus, becomes the task of determining to what extent LLMs can realize a specific manifestation of intelligence: the cognitive capacity to produce meaning. In other words, when authors adopt a mentalist theory to solve The Problem, I claim that they presuppose a cognitive notion of agency by framing it as a case study within the broader ARI thesis.

On the other hand, I argue non-mentalist approaches take a step toward the non-cognitive side by rejecting the idea that the problem should be understood in terms of whether LLMs replicate a cognitive process of meaning production. Instead, non-mentalist theories focus on whether LLMs replicate human modes of linguistic engagement with the environment. Despite this shift, I claim that non-mentalist theories still presuppose a *soft* cognitive notion of AI agency. The fundamental reason is that they continue to frame the question in terms of *whether* LLMs replicate human modes of linguistic engagement. In contrast, I argue that an approach to the problem of meaning grounded in a non-cognitive notion of AI agency, *à la* Floridi, would instead focus on *how* LLMs's outcomes are linguistically engaging with the environment by non-cognitive

means. Thus, I show that no existing non-mentalist approach has addressed The Problem from a non-cognitive perspective of AI agency, highlighting the gap I aim to fill in Chapter 2.

I finish the argument by comparing these two approaches—namely, framing The Problem by adopting mentalist and non-mentalist theories of meaning—and arguing in favor of the latter. In particular, I claim that the latter route is a less theoretically demanding path as it does not impose cognitive constraints regarding LLMs’ linguistic agency.

This argument is structured around three main components, each corresponding to a section of the chapter. In Section 1.2, I provide a comprehensive overview of Floridi’s non-cognitive framework for artificial agency. Specifically, I present the cognitive and non-cognitive conceptions of AI (Section 1.2.1), the historical interpretation of AI development that underpins Floridi’s rejection of the cognitive approach to AI (Section 1.2.2), and Floridi’s final account of non-cognitive agency (Section 1.2.3).

In Section 1.2, the main topic is The Problem. I begin by outlining its roots in the grounding problem and showing how this issue reemerges in a slightly different form concerning LLMs (Section 1.2.1). I then map the existing literature on The Problem and provide a systematic overview of the strategies proposed to address it (Section 1.2.2). I conclude this section by elaborating a technical introduction to the training and architectural features of LLMs, highlighting how approaches to The Problem depend on such features to support their conclusions (Section 1.2.3).

Finally, in Section 1.3, I present the central argument of this chapter. I begin by distinguishing between proposals that address The Problem by adopting mentalist theories of meaning and those that adopt non-mentalist theories (Section 1.3.1). Then, I demonstrate how mentalist theories align with the ARI thesis, presupposing a cognitive notion of AI agency, whereas non-mentalist proposals align with a soft version of this notion. I finally argue in favor of non-mentalist approaches to The Problem due to their lower theoretical burden (Section 1.3.2).

2.2 A Non-cognitive Notion of Artificial Agency

In two recent works, Floridi (2023b, 2025) elaborates a non-cognitive approach to artificial agency. Since its inception, AI has been understood in two contrasting ways: either as a branch of cognitive science focused on the production of intelligence or as a branch of engineering focused on the reproduction of intelligent behavior. These two conceptions have defined different long-term goals and guided the field’s development in distinct directions. Floridi’s approach to artificial agency begins with the adoption of the non-cognitive conception of AI. The author argues in favor of this view by highlighting the success of the non-cognitive, or engineering, perspective over the cognitive one. Although we remain far from achieving Artificial General Intelligence (AGI), which constitutes the ultimate goal of the cognitive approach, the engineering perspective has proven remarkably effective in designing technologies that replace humans in various tasks by integrating smart technologies, also called AI-related applications, across different domains (Floridi, 2023b, p. 20).

Grounded in a rejection of the cognitive perspective on AI, Floridi develops a conceptual reconstruction of the digital revolution. In particular, the authors interpret the history of digital technologies in terms of their cleaving power, that is, their ability to couple, decouple, and recouple the basic blocks of our modern world. This historical process has led to a fundamental

decoupling at the core of AI technologies: that of agency from intelligence. In other words, the cleaving power of digital technologies provides the backdrop against which AI has emerged as a new form of agency. In this conceptual reconstruction, AI agency is defined by its ability to perform tasks that traditionally require intelligence without possessing intelligence itself. As such, AI technologies do not represent a cognitive revolution but a pragmatic one, increasingly taking over tasks that were once the domain of intelligent agents.

In this section, I present Floridi’s non-cognitivist approach to artificial agency by examining three central components: the argument in favor of the non-cognitive over the cognitive conception of AI as a discipline, the thesis of the cleaving power of digital technologies, and the conception of artificial agency as the decoupling of agency from intelligence.

2.2.1 Cognitive and non-cognitive perspectives of AI

AI has developed along two main lines of research: one as a branch of engineering focused on replicating intelligent behavior and the other as a branch of cognitive science focused on generating intelligence. These two perspectives correspond to what Floridi refers to as the *two souls of AI*. The first line of research defines, as a long-term goal, reproducing successful outcomes of human intelligent behavior by non-biological means. On the other hand, the long-term goal of the cognitive research line is to create the non-biological equivalent of human intelligence. In other words, one approach aspires to build truly intelligent technologies, whereas the other focuses on designing systems that solve specific tasks in ways similar to how humans would.

This division dates back to the early days of AI (Floridi, 2023b, p. 22). For example, on the one hand, traces of the engineering approach can be found in one of the earliest definitions of AI, formulated by McCarthy, Minsky, Rochester, and Shannon in their proposal for the Dartmouth Summer Research Project on Artificial Intelligence. In this foundational text, the authors wrote: “the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving” (Floridi, 2023b, p. 16). On the other hand, early expressions of the cognitive perspective can be seen in recurring attempts to build analogies between computers and the human brain. For example, in a BBC radio broadcast, one of the founding figures of the discipline, Alan Turing, stated: “If it is accepted that real brains, as found in animals, and in particular in men, are a sort of machine, it will follow that our digital computer, suitably programmed, will behave like a brain.” (Floridi, 2023b, p. 18).

As noted, the starting point of Floridi’s approach to artificial agency is the rejection of the cognitive perspective on AI. His argument is based on the observation that the most significant breakthroughs in AI are better understood not as steps toward creating truly intelligent systems but rather as the development of technologies that excel at replicating human outcomes within specific domains. Floridi elaborates on this argument through two fundamental examples: Deep Q-network and the progression from AlphaGo to AlphaFold 2. In the first case, the Deep Q-network, a system of software algorithms, achieved impressive performance by learning Atari vintage games based only on screen pixel data and the score it received as feedback. In the second case, the transition from AlphaGo to AlphaFold 2 represented a fundamental shift in perspective: developers moved from trying to win the board game Go to focusing on developing algorithms that help scientists tackle complex problems. This shift in perspective was vindicated in 2020

when AlphaFold 2 solved the long-standing protein folding problem (Floridi, 2023b, p. 22).

In Floridi's view, Deep Q-network and AlphaFold represent remarkable advances in designing systems that produce outcomes indistinguishable from, or even superior to, human performance in complex tasks. But does this imply that the process through which these systems achieve such outcomes is approaching genuine intelligence? According to Floridi, we have no reason to think so. Neither Deep Q-network nor AlphaFold brings us any closer to developing AGI or truly intelligent systems that replicate human understanding, awareness, acumen, sensitivity, concerns, hunches, bio-embodiment, or wisdom (Floridi, 2023b, p. 23). Nonetheless, both Deep Q-network and AlphaFold are undeniably significant achievements in AI. In this respect, these achievements mark a turning point in AI, but only when examined from a non-cognitive perspective of the discipline:

AI will undoubtedly lead to more discoveries and potentially remarkable breakthroughs, especially in the hands of brilliant and thoughtful people. It will also support the management and control of increasingly complex systems. Yet all these spectacular developments may be more easily achieved if a misunderstanding is removed. We have seen that successful AI is not about producing human intelligence but replacing it. A dishwasher does not clean the dishes as I do but, at the end of the process, its clean plates are indistinguishable from mine. Indeed, they may be cleaner (effectiveness) while using fewer resources (efficiency). The same applies to AI. AlphaGo did not play like the Chinese Go grandmaster and world number one Ke Jie, but it won anyway. Likewise, autonomous cars are not cars driven by humanoid robots (androids) seated behind the wheel instead of you; they are ways of reinventing the car and its environment altogether [...]. In AI, the outcome matters, not whether the agent or its behaviour is intelligent. Thus, AI is not about reproducing any kind of biological intelligence. It is about doing without it. Current machines have the intelligence of a toaster, and we really do not have much of a clue about how to move from there (Floridi, Taddeo, and Turilli 2009). That is the fantastic achievement of reproductive AI, which is the successful continuation of human intelligence by other means (to paraphrase Carl von Clausewitz). (Floridi, 2023b, p. 23)

In this fundamental paragraph, Floridi emphasizes that focusing on the process behind AI outcomes and how similar it is to human cognition is a misguided strategy. It represents a misunderstanding of both the technology and the task it solves, much like expecting a dishwasher to clean dishes using the same method as a person or expecting electric cars to be driven by humanoid robots. Floridi rejects this perspective and, in response, adopts the non-cognitive approach. In this approach, the key consideration is how well the outcome replicates human behavior while addressing the relevant task. In this sense, Floridi concludes that the non-cognitive perspective provides a more accurate framework for both understanding AI achievements and guiding the long-term goals of the discipline.

2.2.2 The cleaving power of AI

The Cleaving Power of AI arises as an attempt to conceptualize the transformations brought about by AI technologies. This thesis states that digital technologies cut and paste our reality, and, in doing so, they couple, decouple, and recouple the basic blocks of our modern world. To better understand this transformative capacity, let us consider a few illustrative examples. In the first place, the digital revolution introduced an unprecedented capacity to store information about a person. This information has become increasingly integral to a person's identity, to the point that, in EU legislation, personal data protection is considered a matter of human dignity (Floridi, 2023b, p. 6). In this context, a significant case of coupling occurs between self-identity and personal data. Another relevant example arises when we notice that, in contrast to the generation before the digital revolution, we can now be digitally present in places where we are not physically located. Floridi interprets this as a case of decoupling between location and presence. Furthermore, a relevant example of coupling, better understood as recoupling, occurs between producers and consumers. The original coupling between consumers and producers during hunter-gatherer societies was decoupled during agrarian societies. This decoupling has recently started to be recoupled through digital technologies that allow the same population to produce and consume information, a phenomenon particularly evident in the exchange of information on digital platforms. In this line, the non-cognitive notion of AI agency elaborated by Floridi is understood in terms of one of the fundamental decouples that these I technologies introduce: that one between agency and intelligence.

2.2.3 Decoupling agency from intelligence

Floridi's conceptualization of Artificial Agency is rooted in two key ideas: a rejection of the cognitive perspective of AI and the thesis that AI technologies possess a distinctive cleaving power. When conceptualizing the type of agency that artificial systems introduce, two main paths emerge, depending on whether the cognitive or non-cognitive perspective of AI is adopted. The approach that results from adopting the cognitive perspective begins by expanding the concept of intelligence to encompass AI technologies within its scope. This path involves analyzing whether AI systems can be considered intelligent, if not, why, and if so, determining what type of intelligence they exhibit or could develop. In other words, it involves examining whether intelligence can be replicated or realized artificially, a proposal known as the Artificial Reasability of Intelligence (ARI) Thesis (Floridi, 2025).

On the other hand, an alternative path that Floridi proposes as a consequence of the rejection of the cognitive perspective of AI involves expanding the notion of agency to include non-intelligent forms. Such expansion is grounded in the described historical interpretation of AI technologies and, in particular, in the Claevinig Power Thesis. More specifically, the digital revolution provides the foundations upon which AI technologies transform the environment by enabling processes of coupling, decoupling, and recoupling the basic blocks of the modern world. Within this framework, Floridi claims that artificial agency should be understood as the outcome of a fundamental decoupling: agency and intelligence. In other words, AI technologies introduce a new type of agency, one that does not rely on intelligence as a defining feature. This perspective underpins what Floridi calls the Artificial Realisability of Agency (ARA) Thesis, or

more broadly, the Multiple Realisability of Agency Thesis (MRA), namely, the idea that agency can be instantiated in different forms, including in non-intelligent systems (Floridi, 2025).

In recent work, Floridi (2025) advances a version of the MRA thesis, arguing that artificial agency is merely one among many forms of agency, characterized especially by its lack of intelligence. Within this framework, intelligence is broadly defined as a biological process that requires human cognition, intention, or mental states. Floridi further adds that intelligence involves the ability to transcend predefined objectives through self-determination (Floridi, 2025, p. 18). Concerning the notion of agency, Floridi (2023b) adopts a minimalist approach from the literature on multi-agent systems (Weiss, 2013; Wooldridge, 2009), in which the concept is defined in terms of:

- (a) receiving and using data from the environment through sensors or other forms of data input,
- (b) taking actions based on the input data autonomously to achieve goals, through actuators or other forms of output, and
- (c) improving performance by learning from interactions (Floridi, 2023b, p. 10).

AI technologies, thus, are a novel form of agency that exhibits intelligent behavior without depending on human cognition, intentions, mental states, or the capacity to pursue non-predetermined goals. Although AI systems remain constrained by their predefined operational frameworks, they nonetheless qualify as agents. These new technologies, thus, represent a divorce between agency and intelligence, a new kind of artificial agency that, unlike its human counterpart, lacks cognition.

Artificial Agency is fundamentally constrained by its design and operational framework. Their heavy dependence on training data and predefined objectives shapes their agency in ways fundamentally different from biological systems. The absence of consciousness, intelligence, and understanding creates an unbridgeable gap between artificial and biological-human agency. This limitation extends to the inability to generate truly original, chosen, or preferred purposes or goals, including the ability to choose whether to choose, confining artificial agents to operate within pre-established frameworks. The boundary of "understanding" remains firmly at pattern recognition and matching, without crossing into genuine comprehension. (Floridi, 2025, p. 19).

It is worth mentioning that Floridi does not directly argue against skeptical views that deny the possibility of attributing agency to AI systems. Instead, the argumentative strategy consists of making explicit two conceptual paths regarding agency in AI technologies: one aligned with the ARI thesis, which stems from accepting a cognitive approach to AI, and the other represented by the MRA, more precisely by the ARA thesis, which arises from rejecting the cognitive approach.

Floridi's argument focuses on endorsing the non-cognitive perspective of AI and the corresponding ARA thesis over its cognitive counterpart. The author supports this position by appealing to scientific evidence, Ockham's razor, and an expanding body of scholarly literature. Fundamentally, the claim is that between ARA and the ARI thesis, the latter is more parsimonious, as it avoids cognitive assumptions, and is better supported by AI developments, given that, while non-cognitive AI has already surpassed its goals, cognitive AI is still far from replicating human intelligence. In this sense, Floridi does not aim to disprove the attribution of agency to AI *per se*,

but rather to advocate for a non-cognitive interpretation of agency as a more parsimonious and plausible conceptual framework. In other words, framing AI agency in terms of the ARA thesis constitutes a more reasonable path with a lower theoretical burden.

Furthermore, according to Floridi, cognitive approaches to AI agency risk misallocating efforts and resources toward an unrealistic and speculative ambition: replicating human intelligence. In contrast, by adopting a non-cognitive approach, we can avoid anthropomorphic fallacies and maintain reasonable expectations regarding the capabilities and limitations of these technologies. These considerations serve as a key non-theoretical argument favoring engineering perspective of AI:

This shift in perspective is not merely theoretical. As AI systems become increasingly embedded in various aspects of human society, recognizing their fundamental nature as unintelligent agents—rather than intelligent systems (which may also act as agents)—is critical for their effective design, development, deployment, governance, and regulation, both ethically and legally. Furthermore, understanding AI as Artificial Agency offers valuable insights into its potential and limitations, and helps develop a more suitable framework for its development and integration into human social structures and interactions. (Floridi, 2025, p. 20).

Adopting a more straightforward explanation of AI's capacities and limitations, that is, the MRA or ARA thesis, constitutes the most productive standpoint for designing technologies that contribute to the development of better societies. In other words, a non-cognitive perspective on AI offers a more effective framework for examining the transformative power of AI technologies and guiding their impact in developing more just, diverse, and sustainable societies. With a better understanding of the first debate underpinning this chapter, AI agency, I now pass to the fundamental second debate, The Problem of Meaning in LLMs.

2.3 The Problem of Meaning in LLMs

In recent years, Large Language Models (LLMs) and other forms of Generative AI models, capable of producing linguistic, visual, audio, and different types of content, have attracted significant attention in the field of philosophy of AI (Cappelen & Dever, 2021; Floridi, 2023a). One of the fundamental questions addressed in the literature is whether the outcomes of LLMs can be regarded as meaningful (Bender & Koller, 2020; Grindrod, 2024; Mandelkern & Linzen, 2024). The historical roots of this problem can be traced back to the Symbol Grounding Problem, formulated by Harnad (1990) within the paradigm of symbolic AI. Within this paradigm, AI systems are designed to manipulate symbols through logical rules. Once it is noticed that a symbol must be somehow linked, grounded, or "hooked onto the world" to possess meaning, the challenge becomes explaining how such meaning could arise through nothing more than logical manipulation.

With the rise of statistical AI, this classic problem resurfaced in a new form. Unlike symbolic systems, LLMs are a product of the statistical paradigm, relying on artificial neural networks trained on vast amounts of data to encode and process information. Although the underlying mechanisms between symbolic and statistical AI systems differ, the philosophical concern returns

in the form of the question: (1) *Do LLMs produce meaningful outcomes?* (Bender & Koller, 2020; Grindrod, 2024; Mandelkern & Linzen, 2024; Mollo & Millière, 2023; Piantadosi & Hill, 2022; Piantadosi et al., 2024).

This initial question has been framed in the literature as the tension between two distinct sets of concepts. The earliest papers on the topic (Bender & Koller, 2020; Bender et al., 2021), along with the subsequent responses it inspired (Grindrod, 2024; Lederman & Mahowald, 2024; Mandelkern & Linzen, 2024; Mollo & Millière, 2023; Piantadosi & Hill, 2022; Piantadosi et al., 2024), present question (1) a tension between meaning, reference, and LLMs. This line of thought leads to a more specific version of the question (1), namely, (1.a) *Do the symbols that compose LLMs' output sequences refer in the way other linguistic expressions are traditionally understood to refer?* More recently, however, the literature has begun to explore a tension between a different set of concepts, namely, meaning, communication, and LLMs. This leads to a closely related, though differently formulated, question: (1.b) *Can LLMs engage in meaningful communicative interactions?* (Bottazzi Grifoni & Ferrario, 2025; Gubelmann, 2024a, 2024b; Pepp, 2025). From now on, I refer to question (1) and its two specific reformulations, (1.a) and (1.b), as The Problem.

Several solutions have been proposed in the philosophical and computational linguistics literature to address The Problem. Despite their differences, all these solutions tend to follow the same general strategy, which makes it possible to map them systematically. However, to the best of my knowledge, no prior work has systematically mapped these solutions. In this section, I aim to fill that gap by developing a three-step framework that outlines the common strategy underlying these approaches.

In this section, I provide a detailed study of The Problem, focusing in particular on explaining this three-step framework. In the first section, I briefly elaborate on the historical roots of The Problem, beginning with its formulation in the context of symbolic AI and following its reemergence within the paradigm of statistical AI. I then, through a comprehensive map of the literature of The Problem, elaborate on the three-step framework. Finally, I offer an overview of the two key technical features on which the three-step framework normally relies, namely, their training and architectural features.

2.3.1 The roots of the problem of meaning in LLMs

The grounding problem, initially formulated by Harnad (1990) within the paradigm of Symbolic AI, refers to the difficulty of explaining how symbols in an AI system, which are only logically manipulated, can acquire meaningful content. The challenge arises from the intuition that meaning seems to require a link or grounding in the world, something that symbolic systems seem to lack.

Motivated by the well-known "Chinese Room" thought experiment J. Searle (1980), Harnad (1990) introduces the Chinese/Chinese Dictionary-Go-Round argument through the following thought experiment: suppose Sara is a philology student, trying to learn Chinese using only a Chinese/Chinese dictionary. How successful could she be in understanding the language just by relying on this dictionary? Although a very challenging task, her success would strongly depend on her prior knowledge of other languages and real-world experiences, similar to cryptographers deciphering ancient languages or secret codes by relying on external context and prior linguistic

understanding. In an initial attempt to understand the challenge that symbolic AI systems face, one might be tempted to parallel it with Sara, or the cryptographers, as these systems are essentially trying to learn a new language only through the logical manipulation of symbols.

However, Harnad points out that the situation is even more challenging for symbolic AI systems. Fundamentally, Sara and cryptographers enjoy an extra tool that symbolic AI systems lack, namely, their prior knowledge of other languages and the world. In this sense, a more accurate analogy of the challenge that faces symbolic AI systems would be if Sara were attempting to learn Chinese as their first language using only a Chinese/Chinese dictionary or if cryptographers were deciphering an ancient language without prior knowledge of any other language or the world. Through this analogy, Harnad illustrates that, without prior grounding, symbolic AI models are trapped in a cycle of symbols referring exclusively to other symbols. This leads to the central question of the symbol grounding problem: How can an AI model escape this symbol-to-symbol cycle? In other words, how can a symbol’s meaning be grounded in something that escapes a network of meaningless symbols? (Harnad, 1990, p. 340).

Recently, the philosophical concern surrounding the grounding problem has reemerged in the literature on philosophy of AI (Bottazzi Grifoni & Ferrario, 2025; Floridi, 2023a; Grindrod, 2024; Gubelmann, 2024a; Pepp, 2025) and computational linguistics (Bender & Koller, 2020; Gubelmann, 2024b; Lederman & Mahowald, 2024; Mandelkern & Linzen, 2024; Piantadosi et al., 2024). This challenge, however, takes a distinct form when applied to LLMs, as these models operate within the statistical paradigm rather than the symbolic AI paradigm. LLMs, as deep neural networks with specific training and characteristic features, do not depend on logical manipulation of symbols, but rather on next-token prediction. More specifically, in the literature the relevant question is whether LLMs can generate meaningful outcomes by predicting the next words based on two fundamental features: being trained on a vast amount of text, essential to refine the model’s prediction, and a specific neural network architecture that encodes and processes linguistic patterns, such as word dependencies or world relationships (Bender & Koller, 2020; Grindrod, 2024; Mandelkern & Linzen, 2024).

Despite the technical differences between symbolic and statistical AI systems, the philosophical concern underpinning the grounding problem persists. Fundamentally, the classic grounding problem has reemerged in the literature in two distinct forms: the tension between LLMs, meaning, and reference, on the one hand, and the tension between LLMs, meaning, and communication, on the other. Consider any given expression ϕ , which may occur as a spoken, written, or thought content produced by human agents, and as a part of an output sequence generated by an LLM. For simplicity, from now on, $HA(\phi)$ represents instances of ϕ as produced by a human agent, and $AA(\phi)$ denotes those generated by an artificial agent, such as LLMs.

The first tension arises from considering whether the explanation of the meaning of $HA(\phi)$ can be extended to account for the meaning of $AA(\phi)$. Roughly speaking, theories of reference explain the meaning of ϕ in terms of its reference. Human agents are in a favorable epistemic position to establish a causal or otherwise connection with the reference of ϕ . This connection, thus, is central to explaining the capacity of human agents to understand and produce meaningful expressions. The tension arises when it is noticed that LLMs don’t seem to be in a good position to establish a causal, or any other kind of connection, with the reference of ϕ . Thus, the standard

explanation of meaning for $HA(\phi)$ may not apply to $AA(\phi)$. This raises the first question that underpins The Problem: (1.a) *Do the symbols that compose LLMs' output sequences refer in the way other linguistic expressions are traditionally understood to refer?* (Bender & Koller, 2020; Grindrod, 2024; Mandelkern & Linzen, 2024; Mollo & Millière, 2023; Piantadosi & Hill, 2022; Piantadosi et al., 2024)

The second tension arises from considering whether the condition under which $HA(\phi)$ successfully communicates something can also account for $AA(\phi)$. When considering human agents, such conditions are often described in terms of a particular mode of engagement with the linguistic environment. The tension arises when one asks whether LLMs are capable of engaging with the linguistic environment similarly to how humans do. Thus, without proper justification, the condition under which $HA(\phi)$ communicates something cannot be extended to $AA(\phi)$. This leads to the second question for The Problem: (1.b) *Can LLMs engage in meaningful communicative interactions?* (Bottazzi Grifoni & Ferrario, 2025; Gubelmann, 2024a, 2024b; Pepp, 2025).

2.3.2 Addressing the Problem of Meaning in LLMs

In what follows, I elaborate on a novel three-step framework that outlines the common strategy underlying the different solutions proposed in the computational linguistics and philosophical literature to solve The Problem.

Step 1: Adopting a theory of meaning

The first step is to adopt one of our best philosophical theories of meaning. More specifically, the authors typically begin by adopting one of the well-established theses within various theories of meaning belonging to both philosophy of language and philosophy of mind. Different theories of meaning start their conceptualization by assuming different bearers of meaning, ranging from proper names (Kripke et al., 1980; Putnam, 1975) to sentences (Brandom, 1994) or speech acts (J. R. Searle (1969)). In consequence, these theories encounter different challenges. In some cases, the central challenge is how to fix the meaning of a proper name; in others, it is how to explain the meaning of a sentence or define the condition of a success of a speech act. I recognize that the attempt to summarize approaches to The Problem based on such a variety of theories of meaning carries a risk: sacrificing precision for the sake of generality. Yet, I claim this trade-off is justified as this reconstruction captures the overarching strategy, underpinning all these approaches. Thus, the readers should be aware that in what follows, I will at times introduce certain ambiguities for the sake of generality. For example, I will use ϕ to stand for a proper name, a sentence, or a symbol as needed.

What follows is an enumeration of some of the main theses adopted in the literature addressing The Problem. For each case, begin by citing the relevant work, then present the corresponding thesis, and finally, identify, if applicable, the pioneering theory of meaning that the authors cite as underpinning the relevant thesis. It is worth mentioning that I do not endorse any of the following theses; this is merely a faithful summary of the positions advanced by the respective authors.

- (1.1) Bender and Koller, 2020 (p. 5187): *The meaning of a symbol ϕ is fixed by the entity in the world to which ϕ refers*, no pioneering philosophical theory of meaning cited.
- (1.2) Piantadosi and Hill, 2022 (p. 846): *The meaning of a symbol ϕ is fixed by the conceptual relation that the internal mental representation of ϕ has with the other internal mental representations*; pioneering philosophical theory of meaning cited: conceptual role semantics as developed by Block (1998).
- (1.3) Mandelkern and Linzen, 2024 (p. 1192): *The meaning of a word ϕ is fixed by the history of usage associated to ϕ in the linguistic community*, pioneering philosophical theory of meaning cited: externalist theory as developed by Putnam (1975).
- (1.4) Mollo and Millière, 2023 (p. 2): *The meaning of symbols ϕ is explained¹ through a grounded internal states in which the speaker mentally represents ϕ* ; although no philosophical theory of meaning is cited, the authors reference as contemporary proponents of this thesis Millikan (2017), Neander (2017), and Shea (2018).
- (1.5) Gubelmann, 2024b (p. 11669): *The meaning of a sentence ϕ is explained by the pragmatic norms that regulate its use in a communicative context*; pioneering philosophical theory of meaning cited Brandom (1994, 2010, 2021) and Hlobil and Brandom (2024).
- (1.6) Gubelmann, 2024a (p. 9): *The meaning of a speech act ϕ is partially explained by the speaker's capacity for action*; pioneering philosophical theory of meaning cited Austin (1962) and J. R. Searle (1969)²
- (1.7) Grindrod, 2024 (p. 16): *The meaning of a name ϕ is determined by the naming practice regarding ϕ that is active within the linguistic community*; pioneering philosophical theory of meaning cited Evans (1982).
- (1.8) Grindrod, 2024 (p. 19): *The meaning of a linguistic token ϕ is fixed by the reproductive history of ϕ* ; pioneering philosophical theory of meaning cited Millikan (1984, 2004, 2005).
- (1.9) Bottazzi Grifoni and Ferrario, 2025 (p. 15): *The meaning of an expression ϕ consists of its consistent use under public agreement, shared judgments, and definitions within situated language games.*; pioneering philosophical theory of meaning cited Wittgenstein (1953).

Step 2: translating the thesis regarding meaning to LLMs

The second step that authors follow when addressing The Problem involves inferring, from the adopted thesis of meaning, under which conditions an expression, a word, a sentence, a speech act, a name, and so on, is considered to successfully bear meaning. I suggest calling Condition (2). In general, moving from Thesis (1) to Condition (2) often requires extending the pioneering theory of meaning initially adopted. At this stage, authors frequently diverge considerably from the original theory.

¹In some cases, the reader will notice that I use the term "fixed", while in others I use "explained". This distinction depends on whether the authors take the pioneering theory they adopt to be focused on fixing reference, or rather on providing an explanation of meaning that does not necessarily involve reference-fixing.

²The reader will notice several instances in which the author(s) adopting a given thesis to address The Problem take considerable distance from the pioneering philosophical theory of meaning they cite. For example, in this case, the speaker's capacity for action is not a relevant aspect in Searle's and Austin's work. The reason for this distance is that the author focuses on a specific feature—in this case, speech acts as an action—and uses it as the key to solving the Problem of Meaning. Thus, it is worth mentioning that some of these theses are not strongly grounded in the pioneering theory of meaning but are rather inspired by it to develop a novel approach to solve The Problem.

A positive answer to The Problem depends on whether LLMs satisfy Condition (2). What follows is an overview of these conditions, each inferred from the corresponding thesis enumerated in the previous section. In particular, each Condition (2) corresponds to the same enumerated Thesis (1) presented above.

$AA(\phi)$ successfully bears meaning if it is grounded in:

- (2.1) a causal contact with the object that constitutes the reference of the word ϕ , or
- (2.2) an internal mental representation that captures the conceptual role of the expression ϕ , or
- (2.3) the history of the meaningful usage of ϕ as determined by the linguistic community, or
- (2.4) an internal mental representation of ϕ properly linked to the world, or
- (2.5) the pragmatic norms that govern the use of ϕ , or
- (2.6) capacity for action, in particular, capacity to perform the speech act ϕ ,
- (2.7) the intention to engage in the naming practice associated with ϕ in the speaker's linguistic community, or
- (2.8) the right reproductive history of the linguistic token ϕ ,
- (2.9) constancy in maintaining reference points of ϕ through agreement in both definitions and judgments within situated language games.

Rooted in these conditions, authors frame The Problem as a question of whether LLMs satisfy the relevant condition. Questions that can be found in literature are: *Are LLMs' outcomes properly grounded in causal relationships with elements in the world* (Bender & Koller, 2020), *Are LLMs' outcomes grounded in internal representations that capture the conceptual role of an expression?* (Piantadosi & Hill, 2022; Piantadosi et al., 2024), or *Are LLMs' outcomes grounded in the pragmatic norms that govern the use of the expression ϕ* (Gubelmann, 2024b).

Step 3: Applying the framework normatively

To answer the relevant question behind every specific formulation of The Problem, the final steps involve applying the theory of meaning as a normative framework to investigate the technical features of LLMs. The normativity attributed to these semantic theories becomes evident when it is noticed that the adopted theses (1) along with condition (2) are interpreted as normative principles that determine the facts that would justify a positive or negative answer to the Problem. Let me clarify this point: meaning can be normative in different ways (Boghossian, 2003; Glüer & Wikforss, 2009; Glüer, Wikforss, & Ganapini, 2009). Fundamentally, meaning can be normative either because it *engenders* normativity or because it *determines* normativity (Glüer & Wikforss, 2009; Glüer, Wikforss, & Ganapini, 2009). Roughly speaking, in the first case, a principle concerning meaning holds because certain things have a given meaning. In the second case, certain things have a given meaning because a principle concerning meaning holds. The normativity I attribute to thesis (1) and the condition (2) is the second kind: LLMs outcomes are meaningful or (or meaningless) because (1) and (2) hold. Thus, to address the Problem, the strategy involves arguing that, based on specific technical characteristics of LLMs, one can determine whether the relevant condition is met, and, consequently, whether the model produces meaningful outcomes. Fundamentally, these proposals typically rely either on the architecture or

training features of LLMs. In the next two sections I elaborate on this point by introducing these two features and explaining how they have been used to respond to The Problem.

2.3.3 A technical introduction to LLMs:

LLMs' architecture

Firstly, to better understand proposals that rely on the architecture of LLMs, I provide a step-by-step overview of how these models process input data, starting from the *input embedding layer*, moving through the *positional embedding*, and finishing in the *transformer layer*. This overview is based on the work of Grindrod (2024), Wang et al. (2024), and Zhao et al. (2024), and it takes GPT-3 as an illustrative example.

First, GPT-3 does not process words directly but operates on numerical inputs. Thus, an initial input sequence of symbols $X = (x_1, \dots, x_n)$ is tokenized, that is, it is broken down into smaller units, such as words or subwords, in a format that the model can interpret. Once the sequence is tokenized, each token corresponds to a one-hot vector in which the dimensions correspond to the size of the vocabulary. A one-hot vector is a vector where 1 marks the token's index and all other positions are set to 0. This process can be illustrated with the following toy example. Suppose one is building a minimal model with the following predefined vocabulary:

$$V = \{\text{the, girl, is, now, singing}\}$$

and a maximum input length of three words. In this model, the first step involves defining a vocabulary including all the words and assigning a unique index to all of them as follows:

the = 0, girl = 1, is = 2, now = 3, and singing = 4.

Now, suppose one inputs the sequence “the girl is”. After the sequence is tokenized, the first step involves transforming these tokens into a one-hot vector of five dimensions, based on the size of the vocabulary as follows:

1. the = [1, 0, 0, 0, 0],
2. girl = [0, 1, 0, 0, 0],
3. is = [0, 0, 1, 0, 0].

Thus, the final result of this first step involves encoding the input sequence into a matrix $I_E \in \mathbb{R}^{3 \times 5}$. Real models employ much higher-dimensional representations. In GPT-3, for instance, the vocabulary consists of 50,257 tokens, and the maximum input length is 2,048 tokens (Wang et al., 2024). Each input token corresponds to a one-hot vector of size 50,257. Thus, in real models, the final input sequence is represented in a matrix $I_E \in \mathbb{R}^{2048 \times 50257}$ (Wang et al., 2024).

However, to improve computational efficiency, these sparse vectors, in which most of the elements are 0, are not used directly. Instead, to condense them, each token is transformed into a dense vector via multiplication with an embedding matrix $W_E \in \mathbb{R}^{50257 \times 12288}$. This generates an embedding matrix $X_{\text{WordEmbedding}} = I_E \times W_E$. More specifically, $X_{\text{WordEmbedding}} \in \mathbb{R}^{n \times 12288}$, where $n \leq 2048$ is the number of tokens in the input sequence (Wang et al., 2024).

A fundamental complement to the input embedding layer is the *positional encoding layer*, which encodes information about the order of tokens (Wang et al., 2024). The relevance of this encoding becomes evident when analyzing cases in which a slight alteration in word order

completely changes the meaning of a sentence. For example, consider how modifying the location of the words “excellent” and “mediocre” significantly alters the sentence’s interpretation:

1. “She is an excellent person but a mediocre chef,” and
2. “She is a mediocre person but an excellent chef.”

The positional encoding layer enables the model to capture and preserve meaningful word order, avoiding confusion in cases like (1) and (2) (Wang et al., 2024). After both the input embedding and the positional encoding layers, the information captured by each is intuitively “put together” through *element-wise* addition. This binary operation adds two matrices of the same size to produce a third. This resulting matrix serves as the input matrix $X = X_{\text{WordEmbedding}} + X_{\text{PositionalEncoding}}$.

A third core component of GPT-3 is the *stack of transformer layers* (96 in the largest version). The transformer layer includes a *masked multi-head self-attention* mechanism and a *feedforward neural network* (Wang et al., 2024). *Masked multi-head self-attention* is a fundamental feature for understanding the information these models capture. To explain it intuitively, following Wang et al. (2024), I will begin by introducing *self-attention* mechanisms, then explain why this mechanism is *multi-head*, and finally, why it is *masked*.

First, consider the following two sentences:

3. “The bank opens during the morning.”
4. “They are sitting in the bank feeding the ducks.”

The word *bank* appears in both but has different meanings depending on context. The self-attention mechanism allows the model to disambiguate word meanings based on context. It does so by computing new representations for each token by attending to other relevant tokens in the sequence (Grindrod, 2024). This process occurs within *multi-head self-attention* layers. Multiple attention heads run in parallel. This way, the model captures different types of dependencies or relational features among tokens (Grindrod, 2024). Additionally, the masking features in the multi-head self-attention ensure that, when predicting a token x_i , the model cannot attend to any future tokens x_j where $j > i$. Fundamentally, this preserves the causal structure required for text generation (Grindrod, 2024).

Finally, each transformer block also includes a *feedforward neural network* that applies nonlinear transformations to each token embedding independently. The transformations allow the model to capture complex and non-linear information, improving the model’s expressive capabilities (Wang et al., 2024, p. 9).

The feedforward neural network outputs a vector of logits representing the model’s prediction for the next token at each position in the sequence (Wang et al., 2024). Logits are the unnormalized final scores of the model, before being transformed into a probability distribution. This final step is carried out by the *Softmax function*, which converts the raw scores into a valid probability distribution over all possible next words. Specifically, the Softmax function transforms scores into a distribution where higher values are closer to one and lower values are closer to zero. The token with the highest probability is then selected as the next token, and the process repeats (Wang et al., 2024). In this way, the model generates a sequence token by token.

It is worthwhile to pause here and explain why the transformer architecture represented a fundamental advance in terms of the complexity of the information that language technologies can

encode. Before transformers, the architecture typically used in language technologies was based on *static embeddings* (Grindrod, 2024). These embeddings captured relevant semantic relationships. For example, in Word2Vec, country–capital relationships are geometrically structured in the embedding space (Mikolov et al., 2013). This means that the vector difference between “France” and “Paris” resembles the vector difference between “Germany” and “Berlin” (Mikolov et al., 2013).

Architectures based on static embeddings presented two fundamental problems. First, a sentence is just a sequence of embeddings that bear no relations among themselves. Second, a word, such as “bank” in our earlier example, is represented by the same vector, regardless of whether it appears in sentence (3) or (4). More broadly, the problem with static embeddings is that they cannot capture how the semantic information a given word conveys is also determined by the words surrounding it (Grindrod, 2024).

An initial solution to capture this dependency was the introduction of *recurrent neural networks*, in which the key to capturing dependencies lay in hidden states that process information sequentially. However, these neural networks also presented two problems. First, they cannot effectively capture long-range dependencies in a sequence. For example, consider the following two sentences:

5. “Anna put on her jacket and packed her phone, but she forgot the math book.”
6. “Anna put on her jacket, packed her phone, some gum, the physics book, and the medieval art book, but she forgot the math book.”

A recurrent neural network may successfully capture the dependence between “Anna” and “she” in sentence (5). However, they struggle in capturing the dependence in sentence (6). Recurrent neural networks also presented a second problem: they operate through sequential processing, meaning that each word is processed one at a time. This results in more computationally intensive training.

Transformer architectures provide a solution for all these problems (Grindrod, 2024). Fundamentally, transformer architecture still uses statistical embedding on the first layers, which results from token embedding and positional encoding. However, as remarked, statistical embeddings then pass through a stack of transformer layers composed of both the multi-masked self-attention head and a forward neural network. The result of this is a *contextualized word embedding* that is highly sensitive to many different features of the textual context that the word appears in (emphasis mine Grindrod, 2024, p 6).

Moving back to the Problem, several papers rely on LLMs’ architecture while addressing the Problem. For example, Piantadosi and Hill (2022) and Piantadosi and Hill (2022) start by, following the foundational work of Block (1998), adopting Thesis (1.2): *The meaning of a symbol ϕ is fixed by the conceptual relation that the internal mental representation of ϕ has with the other internal mental representations.* Then, they infer Condition (2.2): *AA(ϕ) successfully bears meaning if it is grounded in an internal mental representation that captures the conceptual role of the expression ϕ .* Then, they frame the Problem as the question *Do LLMs generate outcomes through internal representations that capture the conceptual role of a word?* They interpret internal representations as LLMs’ highly contextualized word embeddings. Finally, the authors argue that contextualized embedding captures such a conceptual role. In more recent work, they go further, suggesting that these contextualized embeddings may even work as a cognitive model

of how humans internally represent concepts (Piantadosi et al., 2024).

Another example of how contextualized word embedding plays a fundamental role in framing the Problem is provided by Mollo and Millière (2023). Accordingly, the start from Thesis (1.4): *The meaning of symbols ϕ is explained through a grounded internal state in which the speaker mentally represents ϕ .* Then, they infer the condition (2.4): *AA(ϕ) successfully bears meaning if it is grounded in an internal mental representation of ϕ properly linked to the world.* According, they frame the Problem as the question *Do LLMs generate outcomes through internal representations grounded adequately into the world.* The notion of internal representation also refers to contextualized embeddings. To give a positive answer to The Problem, the authors argue that it must be shown how these embeddings "hook onto the world" (p. 28 Mollo & Millière, 2023). In particular, they suggest to analyze the training features of LLMs to determine whether these embedding vectors are properly grounded in the world or not.

LLMs' training

Rather than focusing on specific aspects of LLMs' architecture, other authors address The Problem by concentrating on the broader aspect of how these models are trained. There are two distinct training paradigms worth mentioning: traditional fine-tuning and prompting paradigms. On the one hand, the fine-tuning paradigm has been applied to medium-sized models, which typically have up to one billion parameters. On the other hand, the prompting paradigm is designed for models with more than one billion parameters (Zhao et al., 2024). To illustrate the contrast in size between the models associated with these two paradigms, consider that the largest version of GPT-2, representing a medium-sized model, has 1.5 billion parameters and was trained on 40 GB of text data, while the largest version of GPT-3, characterized as a large-sized model, contains 175 billion parameters and was trained on 570 GB of text data (Wang et al., 2024).

In the traditional fine-tuning paradigm, the training process consists of two stages: pre-training and fine-tuning. During the pre-training stage, the model is trained on a vast amount of data. For example, in the case of BERT, this involves approximately 45 terabytes of text data (Zhao et al., 2024). This stage allows the model to learn two fundamental tasks: predicting the next token and identifying general patterns in the data. In the second stage, known as fine-tuning, the model is trained for a specific domain by adjusting several parameters. Broadly speaking, this stage allows the model to contextualize its output for a given domain by fine-tuning aspects that help capture the specific patterns relevant to that domain (Zhao et al., 2024).

The prompting paradigm also consists of two training stages: pre-training and Reinforcement Learning from Human Feedback (RLHF). During the pre-training stage, LLMs are trained on an amount of data that considerably exceeds that used for mid-size pre-trained language models. For instance, in the case of GPT-3, the model utilizes 175 billion parameters and is trained on 570 GB of text data (Zhao et al., 2024). As noted by Wang et al. (2024), unlike pre-trained large language models that traditionally required both pre-training and fine-tuning, LLMs trained under the prompting paradigm do not require a fine-tuning stage. This is because the increase in data during the pre-training stage has shown remarkable performance in enabling the models to effectively achieve specific tasks without additional domain-specific fine-tuning (Wang et al., 2024). However, these models still have two fundamental problems: they cannot follow users' instructions, as

standard chatbots do, and they generate harmful and biased content. RLHF generally solves the first problem, and the second one is solved by instruction tuning via demonstration pairs.

These training features in LLMs have been interpreted in various ways within the literature addressing The Problem. These interpretations begin with two fundamental questions: *What kind of semantic information is embedded in the vast amount of data used to train LLMs?*, and *What kind of semantic information does the model "learn" as it encodes the linguistic patterns present in this data?*

Bender and Koller (2020) begin they answer by distinguishing between the *form* and the *meaning* of an expression. Form refers to any observable realization of language. Meaning, on the other hand, is defined as the relation between the form of an expression and something in the world. Thus, they start from claiming thesis (1.1): *The meaning of a symbol ϕ is fixed by the entity in the world to which ϕ refers*. Then, they infer Condition (2.1): *AA(ϕ) successfully bears meaning if it is grounded in a causal contact with the object that constitutes the reference of the word ϕ* . The clue of their answer lies in interpreting the data used to train LLMs as being the mere *form* of language. Thus, during training, LLMs cannot "learn" meanings; they only encode the *form* of language. This way, the authors provide a negative answer to The Problem: LLMs do not generate meaningful outcomes since, as they are only trained exclusively on the *form* of a language, they fail to meet the key condition for producing meanings, namely, establishing a causal connection with external objects that constitutes the reference of our words.

Proposing a different condition (2) grounded on alternative theories of meaning has become the standard strategy for responding to Bender and Koller (2020)'s negative answer to The Problem (Grindrod, 2024; Gubelmann, 2024b; Pepp, 2025; Piantadosi & Hill, 2022). In these alternative accounts, the central step is to provide a philosophical interpretation of the data used to train LLMs, arguing that it is an entity complex enough to support the kind of semantic information needed to demonstrate that LLMs genuinely produce meaningful outcomes.

For example, Gubelmann, 2024b starts from Thesis (1.5): *The meaning of a sentence ϕ is explained by the pragmatic norms that regulate its use in a communicative context*. From this condition, the authors infer Condition (2.5): *AA(ϕ) successfully bears meaning if it is grounded in the pragmatic norms that govern the use of ϕ* . Then the author frames The Problem as the question: *Is the data used to train LLMs complex enough to capture the pragmatic norms that govern the use of our expressions*. The author suggests that it might be possible to provide a positive answer.

In a similar line, Grindrod (2024), starts by adopting Thesis (1.8): *The meaning of a linguistic token ϕ is fixed by the reproductive history of ϕ* . Then, the author infers condition (2.8): *AA(ϕ) successfully bears meaning if it is grounded in the right reproductive history of the linguistic token ϕ* . Accordingly, the authors frame the problem in terms of the question *Is the data used to train LLMs complex enough to capture the right reproductive history of our expressions*. Grindrod (2024) finishes claiming that it might be possible to provide a positive answer.

2.4 Cognitive and Non-cognitive Approaches to the Problem of Meaning

In the previous section, I presented a systematic investigation of various ways in which The Problem has been addressed in the computational linguistics and philosophical literature. One of the common aspects of these different solutions lies in starting by adopting a theory of meaning. This provides room for characterizing these different solutions based on the adopted theory of meaning. For the present thesis, a relevant characterization emerges from the distinction between mentalist theories, in which theoretical constructs such as mental representations or psychological features of the speakers play an explanatory role in the account of meaning, and non-mentalist theories of meaning, in which these concepts do not play such a role (Speaks, 2010). By adopting this distinction, in the first part of this section, I argue that the various solutions proposed in the literature to address The Problem can be categorized into those that adopt mentalist theories of meaning, on the one hand, and those that adopt non-mentalist theories of meaning, on the other hand.

In the second part of this section, I fulfill the central objective of this chapter: to connect the debate about AI agency with the Problem. In this section, I make that connection explicit. In the first part, I show how approaches to the Problem that rely on a mentalist theory of meaning align with a cognitive perspective on AI. Accordingly, mentalist theories frame the Problem as an instance of the ARI thesis. Under this framework, the Problem is framed as asking whether LLMs produce meaning by replicating the relevant cognitive processes. In contrast, those adopting non-mentalist theories take a step closer to a non-cognitive notion of AI agency. I claim that they occupy a middle ground between cognitive and non-cognitive conceptions of agency. On the one hand, by giving up the idea that the Problem should be understood in terms of whether LLMs produce meaningful outcomes by simulating a process of meaning production, they move closer to a non-cognitive view. On the other hand, by still framing the question in terms of *whether* LLMs' outputs successfully engage in linguistic interaction—rather than *how* they do so—they remain tied to *soft* cognitive notion of AI agency. Fundamentally, by connecting these two debates, I demonstrate that so far, no proposal has been adopted that is a non-mentalist theory of meaning guided by a non-cognitive notion of AI agency as developed by Floridi. In the second chapter, I address this gap.

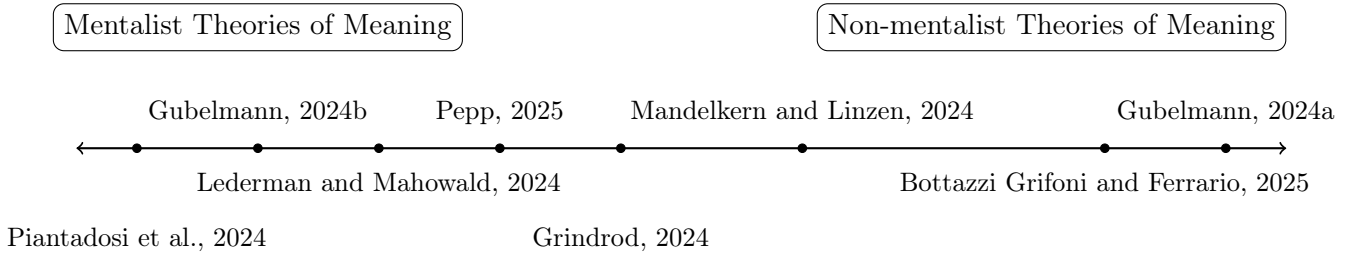
In this final section, I first provide a novel characterization of the solutions addressing The Problem, distinguishing between those that adopt mentalist theories of meaning and those that adopt non-mentalist theories (Section 1.3.1). Finally, show the connection between the debate about AI agency and the discussion about the Problem (Section 1.3.2).

2.4.1 Mentalist and non-mentalist theories of meaning

A considerable range of theories of meaning has been adopted to address The Problem. In this thesis, I distinguish between those that adopt mentalist and those that adopt non-mentalist theories of meaning. Mentalist theories are those that attribute explanatory power to some sort of mental or psychological feature of the speaker while explaining the fact in virtue of which an expression has a given meaning. Non-mentalist theories, on the other hand, reject such

appeal to any such sort of features (Speaks, 2010). Given the complexity of what it is being characterized, namely, theories of meaning, this distinction should not be understood as a strict binary distinction. Instead, it can be used to define a conceptual spectrum for classifying theories according to the extent to which they depend on mentalist assumptions in their explanatory models for meaning.

At the mentalist end of the conceptual spectrum (see Graphic 2.1), I place those proposals that adopt Block (1998)’s conceptual semantics as a foundational work. By doing this, Piantadosi and Hill (2022), Piantadosi et al. (2024) root their proposal in a long-standing tradition of mentalist theories of meaning in the philosophy of mind and cognitive science. Specifically, following Block (1998), the authors begin with the assumption that linguistic expressions inherit their meanings from mental representations. Furthermore, the meaning of such mental representation is determined by its conceptual role. This gives room to the condition (2.2): $HA(\phi)$ is meaningful since the speaker developed a corresponding internal representation that captures the conceptual role of the concept ϕ . They then argue LLMs might satisfy this condition, given that the internal vectors can be interpreted as capturing the conceptual role of a given word. Thus, the meaning of $AA(\phi)$ might be justified on the way LLMs represent words internally.



Graphic 2.1: Conceptual spectrum illustrating the position that occupies different attempts to address The Problem based on the theory of meaning adopted.

Slightly towards the center of the conceptual spectrum, though still on the mentalist side, I place Gubelmann (2024b). In this work, the author follows two foundational texts from the pragmatic tradition in philosophy of language: Austin (1962) and J. Searle (1980). A recurring thesis in the pragmatic tradition is that the meaning of an expression in a given language is explained in terms of the propositional attitudes of the speakers (Grice, 1969). Grounded on the thesis that the meaning of a sentence can be reduced to the speaker’s meaning and, ultimately, to the speaker’s intentions Gubelmann (2024b) develops a negative answer to The Problem. Broadly speaking, the author argues that the condition in virtue of which the speech act $HA(\phi)$ has meaning is that the speaker has the capacity for action, in particular, the capacity to perform $HA(\phi)$. The author continues arguing that LLMs lack the capacity for action; thus, they cannot satisfy (2.6). This way, LLMs do not participate in language in the same way humans do, given that they cannot act, a fundamental condition that enables humans to speak and interact with others.

At the center of the conceptual spectrum, I place proposals that adopt any version of what

has been called in philosophy of language externalist theories of meaning, in which meanings are determined, at least partially, by features of the environment. For example, a fundamental author in this tradition, Kripke (1980), argues that the reference of a name could be explained in terms of its history of usage. Externalist theories of meaning are placed at the centre of the spectrum, as here we can see how internal mental representations begin to lose prominence in terms of their explanatory role. However, many versions of externalism still attribute a secondary explanatory role to speaker intentions. This is especially relevant to understanding how this account explains the meaning of ambiguous expressions, context-sensitive expressions, or any other sort of expression associated with more than one meaning. The standard approach for these expressions involves claiming that they $HA(\phi)$ must be accompanied by a communicative intention of the speaker that will fix which meaning was intended to be expressed (Grindrod, 2024; Pepp, 2025). Similarly, when a term is transmitted from one speaker to another, the successful transition requires that the latter speaker uses the same way as the former (Speaks, 2010). This way, though there is no longer a reduction of meaning to the speaker’s intention, they still play a role in the externalist picture of explaining meaning.

At first glance, externalist theories seem to offer a compelling framework for explaining reference in LLMs. This can be done by simply arguing that the data used to train LLMs encodes the historical usage of names. This interpretation, however, would also require attributing communicative attempts to LLMs. Such a move constitutes a highly controversial claim, as noted by Pepp (2025) and Lederman and Mahowald (2024). In response to this challenge, Lederman and Mahowald (2024) proposes adopting an *interpretationist* perspective from the philosophy of mind. Following this perspective, a system can be said to have intentions if and only if its behavior is well explained by the hypothesis that it does (Davidson, 1973; Davidson & LePore, 1986; Dennett, 1971, 1989). On the other hand, Pepp (2025) addresses this challenge by rejecting the thesis that intentions are required to explain meaning. Following Derosset (2011), Speaks (2016), and Devitt (2021), Pepp (2025) argues that young children can refer, although they have not yet developed the requisite mental sophistication to intend to refer. This suggests that there may be alternative explanations for how LLMs succeed in referring without having the capacity to intend to refer.

While addressing The Problem, Grindrod (2024) also acknowledges the challenges involving the explanatory role that communicative intentions play in the externalist picture of meaning. The author frame it as the challenge of decoupling linguistic intentionality from mental intentionality. Grindrod (2024)’s solution relies on adopting theories of meaning in which linguistic intentionality no longer depends on the speaker’s inner mental states or psychological features, but instead on pre-existing meaningful systems. Solving The Problem, thus, involves investigating whether LLMs can employ this pre-existing system in the same way that ordinary speakers do. Grindrod (2024) adopts two different foundational works to address The Problem in these terms, namely, Evans (1982), on the one hand, and Millikan (1984), Millikan (2004), and Millikan (2005), on the other hand.

Still within the externalist tradition, but moving closer to the non-mentalist end of the spectrum, I place the proposals developed by Mandelkern and Linzen (2024), grounded in the

foundational work of Putnam (1975). Mandelkern and Linzen (2024) adopt from Putnam’s externalism the well-known thesis that the linguistic community fixes meaning. This way, the authors eliminate any dependence on the psychological features of the speakers while explaining the meaning of an expression. As we have seen, grounded on this thesis, the authors then define the following condition (2.3): $HA(\phi)$ is meaningful since the speaker has access to the history of meaningful usage of ϕ as determined by the linguistic community. The authors then suggest a positive answer to The Problem by arguing that it is possible to understand the data used to train LLMs as a complex entity that can carry relevant information regarding the history of meaningful use of our expressions.

At the non-mentalist end of the conceptual spectrum, I place the work of Bottazzi Grifoni and Ferrario (2025), which is grounded in the foundational theory of meaning of Wittgenstein (1953). The authors, in particular, adopt the thesis that successful communication involves maintaining consistency in public agreements on judgments and definitions within situated language games. Based on this thesis, the authors infer that the successful communication of the expression $HA(\phi)$ requires not only regularity in language but fundamental consistency in maintaining reference points through agreement in both definitions and judgments within situated language games. Based on this condition, they develop a negative response to the The Problem’s argument: LLMs lack the consistency needed to track negotiations and contradictions throughout a dialogue. This way, LLMs cannot hold consistency in the reference points necessary for genuine communication. As can be seen, the mental aspects or psychological features of the speaker do not play any explanatory role in this model of successful communication.

Finally, I also place at the non-mentalist end of the conceptual spectrum a solution to The Problem grounded on one of the contemporary figures that continues the tradition initiated by Wittgenstein (1953). Gubelmann (2024a) address The Problem grounded on the work of Brandom (1994), Brandom (2010), Brandom (2021), and Hlobil and Brandom (2024). In particular, Gubelmann (2024a) starts by adopting the thesis that the meaning of a sentence $HA(\phi)$ is understood in terms of the pragmatic norms that regulate its use in communicative contexts. This way, what matters when addressing The Problem is whether LLMs have access to such pragmatic norms that regulate meanings. On this basis, Gubelmann (2024a) provides a positive answer to the The Problem by arguing that the norms observed by language users leave recognizable patterns in the training data, which the LLMs can represent and with them infer the norms governing the use of expressions, that is, the meanings of these same expressions (Gubelmann, 2024a, p. 11670).

2.4.2 A Defense of non-cognitive approaches

In this section, I connect the debate about AI agency, as introduced by Floridi, with the discussions about the Problem. In particular, I show that attempts to address the Problem by adopting a mentalist theory of meaning presuppose a cognitive notion of AI agency. In contrast, those that adopt a non-mentalist theory of meaning are *closer* to endorsing a non-cognitive notion of agency, though still presuppose a *soft* cognitive notion. Furthermore, just as the distinction between mentalist and non-mentalist theories of meaning is not sharp but rather falls along a spectrum, the distinction between presupposing a cognitive or a non-cognitive notion of agency also lies on a

spectrum. Thus, I also show that there is still room to move further toward the non-cognitive end of the conceptual spectrum by endorsing a non-cognitive notion of AI agency. In the next chapter, I address this gap by proposing the first attempt to solve the Problem through the adoption of a non-cognitive notion of AI agency.

The problem of meaning grounded in cognitive notion of AI agency

To generate meaningful outcomes, LLMs must satisfy condition (2), which derives from the adopted theory of meaning. Condition (2) sets the expectation that LLMs must meet to perform effectively, at least from a philosophical standpoint, their primary task: language generation. Accordingly, the role of the chosen theory of meaning is not merely to provide a philosophical account of meaning to address question (1), but also to settle the standard of what counts as ideal technology for language generation. In other words, the theory of meaning adopted serves a dual purpose: determining under what conditions LLMs generate meaningful outcomes, and defining what our expectations are regarding LLMs as a language technology.

But, how exactly does the adopted theory of meaning determine this second dimension in the case of non-mentalist theories of meaning? As noted, approaches that solve The Problem by adopting mentalist theories of meaning define condition (2) in mentalist terms. For example, as argued by Piantadosi and Hill (2022), Piantadosi et al. (2024), and Coelho and Millière (2023), a positive answer to The Problem depends on whether the role of internal vectors in LLMs replicates the role of internal mental representations in human cognition while generating meaningful expressions. Their answer to The Problem, thus, depends on how closely LLMs mimic the cognitive process through which humans generate meaningful expressions. This approach not only provides a path to answer The Problem, but also sets strong expectations for LLMs as language technologies. In short, LLMs are understood as language technologies designed to replicate the cognitive processes by which humans generate the same linguistic outcomes. This expectation emerges from any approach that addresses The Problem by adopting a mentalist theory of meaning. In this sense, mentalist approaches align with Floridi's cognitive perspective on AI, where the ultimate goal is to develop truly intelligent technologies (See Table 2.1).

As noted, within mentalist theories, generating meaningful expression is fundamentally understood as a cognitive process, a specific manifestation of human intelligence. For LLMs to generate the same outcomes, they must replicate this cognitive process as closely as possible. While addressing The Problem from a mentalist approach, the question under investigation is whether one of the particular forms in which intelligence manifests, i.e., production of meaning, can be realized in non-biological artifacts, namely, LLMs. This reveals a clear connection between adopting mentalist theories of meaning and the ARI thesis, namely, adopting a mentalist theory of meaning involves framing The Problem as an instance of the ARI thesis.

Criterion	Cognitive End of the Spectrum	Center of the Spectrum	Non-cognitive End of the Spectrum
(i) <i>Notion of Agency</i>	A cognitive notion: AI technologies instantiate agency requires simulating human-like intelligence.	A <i>soft</i> cognitive notion: AI agency requires simulating human-like linguistic interactions	A non-cognitive notion: AI agency replicates intelligent behavior by non-cognitive means.
(ii) <i>How to Investigate AI Agency</i>	ARI Thesis: Do AI technologies instantiate agency by successfully simulating intelligence?	Do AI technologies instantiate agency by successfully replicating humans' modes of interaction with the environment?	ARA Thesis: How are AI technologies instantiating specific human modes of interaction with the environment by non-cognitive means?
(iii) <i>Theory of Meaning</i>	Mentalist theory of meaning	Non-mentalist theory of meaning	Non-mentalist theory of meaning
(iv) <i>Meaning is fundamentally understood as a...</i>	a cognitive process	a specific mode of engagement with the environment	a specific mode of engagement with the environment
(v) <i>Conceptual Refinement of the Problem</i>	Do LLMs generate meaningful outcomes by replicating the cognitive process underpinning meaning production?	Do LLMs successfully engage with their environment by replicating human modes of linguistic interaction?	How are LLMs linguistically engaging with the environment?
(vi) <i>Key to Addressing the Problem of Meaning in LLMs</i>	Determining whether LLMs replicate the cognitive process of meaning production	Determining whether LLMs replicate human modes of linguistic engagement with the environment	Describing LLMs' actual modes of linguistic engagement with the environment

Table 2.1: Summary of how I connect the discussion about AI agency and The Problem. The first column shows the different concepts and criteria underpinning this connection. Each new criterion, starting from the less general, *Key to Addressing the Problem of Meaning in LLMs*, requires precisely defining the criterion above it; in this case, it requires making explicit the *Conceptual Refinement of the Problem*. The specific form of The Problem, hence, depends on the adopted notion of agency. Additionally, the remaining three columns represent the specific form that this dependence takes. More specifically:

1. The column *Non-cognitive End of the Spectrum* represents the standard mentalist approaches to The Problem, which rely on a cognitive notion of agency.
2. The column *The Center of the Spectrum* represents non-mentalist approaches that come closer to framing The Problem as an instance of the ARA thesis, but still cannot be said to be grounded in a non-cognitive notion of agency *à la* Floridi.
3. The *Non-cognitive End of the Spectrum* illustrates what it would look like to solve The Problem based on such a notion of agency. So far, no proposal follows this route. In the next chapter, I present the first attempt to do so.

The problem of meaning grounded in a *soft* cognitive notion of AI-agency

On the other hand, non-mentalist theories addressing The Problem do not presuppose a *strong* cognitive perspective of AI. While these approaches still establish conditions that LLMs must meet to generate meaningful outcomes, these conditions are no longer tied to an ideal of LLMs as a language technology that must mimic the cognitive process through which humans generate meaningful expressions. Instead, these conditions involve asking whether LLMs encode, for instance, the pragmatic norms, naming practices, or historical use associated with a name within the linguistic community. From a non-mentalist perspective, the central question is whether LLMs encode information and interact with the linguistic environment in a way that allows them to generate meaningful expressions through fundamentally non-cognitive means. Externalist approaches to The Problem illustrate this by avoiding reliance on mental constructs, such as communicative intentions. In particular, as several authors note, a key challenge while addressing The Problem by adopting externalist frameworks is the role that communicative intentions play in fixing reference in different externalist accounts of meaning. This challenge arises from the difficulty of determining whether LLMs possess communicative intentions at all.

However, although non-mentalist theories are closer to instantiating the ARI thesis, they still cannot be characterized as occupying the end of the non-cognitive side of the spectrum of agency, at least not as defined by Floridi. There is a fundamental characteristic of the non-cognitive notion of agency that no non-mentalist approach to the Problem has adopted so far. As remarked, Floridi does not offer an argument against the skeptical stance that AI technologies do not instantiate agency. Instead, he starts from the assumption that AI technologies instantiate a new kind of agency, characterized by their lack of intelligence. Floridi, thus, does not aim to determine *whether* AI technologies instantiate agency, but rather to investigate *how* they instantiate agency by non-cognitive means.

Moving towards LLMs, non-mentalist approaches take the first step toward a non-cognitive approach to the Problem by giving up the question of whether LLMs generate meaningful outcomes through the cognitive production of meaning. However, authors adhering to these non-mentalist theories still frame the Problem in terms of *whether* LLMs successfully participate in linguistic interactions by, for instance, encoding pragmatic norms, histories of use, naming practices, and so on. However, there is still conceptual space to move towards the end of the non-cognitive spectrum. In particular, there is room to frame the Problem of Meaning in terms of *how* LLMs successfully participate in linguistic interactions within the environment.

As I will explain in more detail in the next section, this first step that the non-mentalist proposal undertakes inherits the first advantage identified by Floridi regarding non-cognitivist framework for agency: namely, it constitutes a framework with a lower theoretical burden for exploring how LLMs are being integrated into legal environments. However, it does not secure the second advantage: namely, providing a conceptual framework for understanding the transformation introduced by LLMs—as non-intelligent linguistic agents performing tasks previously carried out by intelligent agents—within the environment. To secure both advantages, we must move towards the end of the continuum spectrum.

To summarize the picture so far, answering the question *Do LLMs generate meaningful outcomes?* requires adopting a theory of meaning. Such a theory not only defines what meaning

is, but also shapes our expectations of LLMs as a language technology. In other words, selecting a theory of meaning also involves presupposing either a cognitive or non-cognitive notion of AI agency. Once the connection between the two debates—AI agency and the Problem—is made explicit, two main paths emerge: understanding meaning in cognitive terms, as a manifestation of intelligence, or understanding it in non-cognitive terms. Mentalist theories follow the first path, asking whether LLMs can artificially instantiate one of the many ways in which intelligence manifests itself, specifically by producing meaning. On the other hand, non-mentalist theories, as adopted so far to solve the Problem, are closer to presupposing a non-cognitive notion of agency. They take a first step in that direction by rejecting the framing of the Problem as the question of whether LLMs simulate a cognitive process of meaning production. However, they cannot be said to occupy the far end of the non-cognitive side of the spectrum of agency, as they still focus on determining the *whether* rather than the *how* of LLMs’ linguistic agency. The end of the spectrum, then, would be occupied by a proposal that frames the Problem of Meaning in terms of *how* LLMs participate in linguistic interactions (see Table 2.1).

A Defense of non-mentalist approaches

Addressing The Problem as an instance of the ARI thesis constitutes a more theoretically demanding route, as it relies on strong cognitive assumptions about agency. This approach requires understanding AI technologies as replicating human intelligence. In particular, it requires understanding LLMs as trying to replicate the cognitive process by which humans generate meaningful expression. In contrast, addressing The Problem in terms of both *soft* cognitive notion of AI agency allows us to study the kind of linguistic agency that these models introduce, but without imposing strict cognitive constraints. Thus, by doing the first step towards a non-cognitive notion of agency, non-mentalist theories enjoy the first advantage that Florida attributed to a non-cognitive framework, namely, lower theoretical burden.

But who could I convince to adopt a non-mentalist theory of meaning, given that it entails a lower theoretical burden for investigating AI agency? This argument challenges those interlocutors who subscribe to mentalist theories of meaning as the best route to address The Problem. However, the extent to which this argument will persuade such interlocutors depends on their underlying motivation for adopting a mentalist theory of meaning. For example, an argument by parsimony will not persuade those who adopt a mentalist theory of meaning as a consequence of a strong commitment to the development of General Artificial Intelligence (AGI). It is unlikely that those who consider AGI as the ultimate goal of AI research and LLMs as a potential step towards that objective might find this argument by parsimony persuasive. This seems to be the case of Piantadosi et al. (2024), and Mollo and Millière (2023) .

The specific target of this argument is those interested in solving The Problem for normative reasons, and who consider that the best route is to adopt a mentalist theory of meaning. This is the case, for example, of the influential series of papers by Bender and Koller (2020) and Bender et al. (2021), where the motivation behind addressing The Problem is to determine how these technologies must be designed and applied. Accordingly, from this perspective, a solution to The Problem lays the groundwork for determining how these technologies should be developed and used.

At first glance, those motivated by normative concerns might view mentalist theories of meaning as offering a more precise framework for assessing the conditions under which we can attribute the capacity to produce meaningful outcomes to LLMs. More specifically, they might argue that adopting strong mentalist constraints makes it easier to develop a critical response to The Problem that highlights how far LLMs remain from achieving genuine AGI. This, in turn, supports a critical stance on the development and deployment of language technologies. The problem with this strategy is that it goes too far in its attempt to develop a critical stance on LLMs' capacity to generate meaningful outcomes. In doing so, it forecloses the possibility of grounding a critical evaluation of language technologies not on the speculative goal of achieving AGI, but rather on a conceptualization of their actual modes of linguistic engagement with the environment. A more parsimonious and plausible route to developing a critical stance on language technologies, therefore, involves taking for granted that LLMs' production of meaning is accomplished by non-cognitive means. This allows us to concentrate our conceptual effort on studying, via non-mentalist theories of meaning, how these models actually engage with the environment.

To summarize the result of this section: addressing The Problem as an instance of the ARA thesis constitutes a compelling route for those motivated by studying the *must* and *must-not* of language technologies. Fundamentally, in contrast with framing The Problem as a case study of the ARI thesis, this approach allows one to remain skeptical about LLMs' capacity to produce meaning in cognitive terms, without abandoning the possibility of understanding and assessing their integration into linguistic environments.

Chapter 3

A Non-cognitive Approach to the Problem of Meaning in LLMs

3.1 Preliminaries

In the previous chapter, I made explicit the connection between two debates: AI agency and The Problem. I showed that while mentalist proposals adopt a cognitive notion of AI agency, non-mentalist proposals rely on a *soft* cognitive notion. In this chapter, I begin by illustrating what a commitment to this *soft* cognitive notion looks like in practice, focusing on Mandelkern and Linzen (2024)'s attempt to solve The Problem by adopting a non-mentalist theory of meaning, namely, Putnam's Externalism. I then reconstruct the decisions that led the authors to presuppose this notion of agency and, in departing from those decisions, I develop a novel approach to The Problem. In doing so, I address the gap identified in the previous chapter by offering the first attempt to solve The Problem grounded in a non-cognitive account of AI agency.

Mandelkern and Linzen (2024)'s commitment to a *soft* cognitive notion of AI agency is made explicit by their framing of the grounding problem in terms of the question of whether LLMs successfully engage with the environment by replicating human modes of linguistic engagement (see Table 3.1). In the first two sections of this chapter, I offer an in-depth reconstruction of the elements that lead the authors to presuppose a soft notion of AI agency. In particular, I show that two fundamental features—when considered together—account for the authors' commitment to such a notion: a semantic methodology and their ambitiously intended scope of application (from now on, ambitious scope). In contrast, guided by a non-cognitive notion of agency, I apply Putnam's theory of meaning by adopting a metasemantic methodology and a modestly intended scope of application (from now on, modest scope). In what follows, I introduce these two features.

To mark the first key difference between my application of Putnam's Externalism and that of Mandelkern and Linzen (2024), I introduce the distinction between semantic and metasemantic theories. In short, semantic theories focus on ascribing meaning to expressions, while metasemantic theories focus on the facts in virtue of which we ascribe meaning to those expressions (Speaks, 2010; Stalnaker, 2017). I argue that both a metasemantic and a semantic methodology are at play in the literature on The Problem. This distinction does not directly hinge on whether one accepts a semantic or metasemantic theory of meaning. As I will demonstrate later in this chapter, some accounts of The Problem adopt metasemantic theories of meaning while still employ

a semantic methodology. Instead, this distinction depends on the expectations placed on the theory of meaning. Semantic methodologies expect the theory to determine the meaning of LLMs' outcomes, while metasemantic methodologies focus on studying the facts that give rise to the ascription of meaning to LLMs' outcomes. In the first section of this chapter, I describe both semantic and metasemantic methodologies and show how the semantic methodology is initially implemented in one of the foundational works on the topic, Bender et al. (2021). I then argue that Mandelkern and Linzen (2024) inherits this methodology in their attempt to respond to Bender et al. (2021)'s paper.

Another key difference between my proposal and Mandelkern and Linzen, 2024's one lies in that they attribute to Putnam's theory of meaning an ambitious scope, while I attribute a modest one. This scope emerges from adopting Putnam's earlier writings on meaning. Specifically, they follow a version of Putnam's theory of meaning in which the challenge of how language "hooks" onto the world is still a valid one. This challenge, central to the analytic tradition in the philosophy of language, focuses on explaining how words can be connected to the world. In his earlier writings on meaning, Putnam addresses this problem. However, in his later work, his focus shifts: rather than directly addressing it, Putnam shows how the thesis that *the linguistic community fixes the meaning of our words* dissolves it. In particular, he claims that this challenge only arises if we overlook the role of the division of labor in fixing meanings within the linguistic community.

Once we move to The Problem, we have two options: on the one hand, one might claim that, if Putnam's Externalism was intended to explain how language hooks into the world, it can then be used to explain how LLMs' outcomes hook into the world as well. I call this first path an *ambitiously intended scope of application* of Putnam Externalism. In contrast, if one accepts that Putnam's theory of meaning actually demonstrates why the challenge of how language hooks into the world is not a genuine issue—and that it only becomes a problem when one overlooks how labor distribution works in the linguistic community—then the application of this theory to solve The Problem in LLMs takes on a slightly different form. From this second interpretation, the external pictures of meaning show us how LLM outcomes are already hooked into the world or, more generally, how LLMs are being integrated into the linguistic community. This is what I call a *modest intended scope of application* of Putnam's Externalism.

I argue that these two aspects in conjunction, namely, a semantic methodology along with an ambitious scope of application of Putnam's Externalism, lead Mandelkern and Linzen (2024) to adopt a version of The Problem that presupposes *soft* cognitive notion of AI agency. Their argument can be summarized as follows: given their semantic methodology, they start from the general question of whether LLMs' outcomes mean something. Then, because of their ambitious scope of application, this question is conceptually refined as the question of how these outcomes hook into the world. Accordingly, the authors claim that LLMs' outcomes would mean something if these outcomes were hooked into the world in a similar manner as speakers' words are hooked into the world. In other words, and following the structure developed in Table (2.1), they frame the problem as the question of *whether* LLMs successfully engage with the environment (in this case, the linguistic community) by replicating human modes of linguistic engagement (in this case, grounding the use of a word in its history of use in the linguistic community)? (see Table 3.1).

This way of framing the problem, as remarked in the last Chapter, presupposes a *soft* cognitive notion of AI agency.

In contrast, I aim to develop an application of Putnam’s theory of meaning no longer grounded in a cognitive notion of AI agency. To do this, I make a fundamental departure from Mandelkern and Linzen (2024)’s proposal by adopting a metasemantic rather than a semantic methodology and by embracing Putnam’s rejection of the challenge of how language relates to the world. This departure leads to a very different application of Putnam’s theory of meaning to address The Problem. First, I do not start from the question of whether LLMs mean something. I directly focus on exploring the facts in virtue of which these outcomes mean something (rejection of the semantic feature). I take for granted that they mean something. Additionally, I do not expect to solve how LLMs hook into the world. Instead, the external picture of meaning shows that LLMs’ outputs are already hooked into the world (rejection of the ambitious scope of application). The departure from these two features led me to frame the problem of meaning in terms of *how* LLMs are (linguistically) engaging with the environment (in this case, the linguistic community). This way of framing the problem, as remarked, presupposes a non-cognitive notion of AI agency (See Table 3.1).

I conclude the argument presented in this chapter by showing how conceptualizing The Problem as an investigation into how LLMs are being integrated into the linguistic community provides a conceptual space to reconcile our two fundamental intuitions. In other words, this conceptualization of The Problem offers a framework for understanding LLMs as artificial speakers (i.e., members of the linguistic community) despite lacking the capacity to simulate a cognitive process of meaning production.

This argument is structured around three main components, each corresponding to a section of the chapter. In Section 3.2, I introduce the distinction between semantic and metasemantic methodologies. First, I begin by introducing the distinction between semantic and metasemantic theories of meaning, outlining their object of study and scope (3.2.1). Then, grounded in this distinction, I define both a semantic and metasemantic methodology to address The Problem (3.2.2). The next step involves exemplifying semantic methodologies by reconstructing the first paper in which I recognize such a proposal, namely, Bender et al. (2021). I then show how Mandelkern and Linzen (2024), in their attempt to answer Bender et al. (2021), inherit the same semantic methodology (Section 3.2.3).

In Section 3.3, I define what constitutes an ambitiously intended scope of application of Putnam’s Externalism to solve The Problem, and I distinguish it from a modest scope of application. To do this, I first develop the Twin Earth Thought Experiment, as proposed by Putnam (1975) (Section 3.3.1). This Thought Experiment underpins what has been understood in the philosophical literature as the Putnamian version of the externalist picture of meaning, with the fundamental thesis being the slogan that *the linguistic community fixes the meaning of our words*. I then briefly review some of the interpretations these Thought Experiments have received in the literature (Section 3.3.2). The next step consists of providing an overview of Putnam’s stance in his later writings, where the picture of meaning outlined by the Twin Earth Thought Experiment does not solve how language hooks into the world (Section 3.3.3). I finish this section describing my modest scope application of Putnam’s Externalism in contrast with

the ambitious scope defended by Mandelkern and Linzen (2024).

Finally, in Section 3.4, I begin by making explicit the scope of Putnam’s Externalism in conceptualizing The Problem from both a semantic methodology and a modest application. I then apply this theory along with the two dimensions of meaning defined by it: extension, as determined by the linguistic community (Section 3.4.1), and individual capabilities, as determined by the role of a speaker in the linguistic community (Section 3.4.2). I conclude this section by showing how this application defines a conceptual space that reconciles the two fundamental intuitions underpinning the main goal of this thesis (Section 3.4.3).

3.2 Semantic and Metasemantic Methodologies

3.2.1 Semantic and metasemantic theories of meaning

The distinction between semantic and metasemantic theories of meaning was first explicitly formulated by Speaks (2010) and Stalnaker (2017), but its roots can be traced back to the contributions of Kaplan (1989) and Lewis (1970). In *General Semantics*, Lewis (1970) distinguishes between two kinds of tasks: one involves describing abstract semantic systems that define the relation between symbols and aspects of the world; the other involves describing the psychological and sociological facts that explain why a particular abstract semantic system is adopted by a person or population (Lewis, 1970, p. 19).

Kaplan (1989) further elaborates on this distinction in *Afterthoughts*, by distinguishing types of facts that answer either semantic or metasemantic questions. On the one hand, explaining *that* Φ means α involves studying semantic facts about a given language. On the other hand, explaining *why* Φ means α involves studying historical or sociological facts about that language. While the first type of explanation constitutes a semantic investigation of a language, the second represents a metasemantic or foundational approach to a language (Kaplan, 1989, pp 573-74).

Similarly, Stalnaker (2017) introduces a distinction between descriptive and foundational semantics. On the one hand, descriptive semantics primarily focuses on answering the question of what the semantic value of a given expression is. These theories typically begin by defining a minimal unit of analysis—such as a word or an entire sentence—and assigning meaning to that unit. From there, the meaning of more complex expressions is derived based on the meanings assigned to the minimal units. A fundamental limitation of these theories is that they do not determine the kinds of facts that justify the descriptive theory itself. This second task, on the other hand, falls under the objectives of foundational semantics. These theories explore the nature of facts that explain why a given descriptive semantics is adopted. These facts may be behavioral, mental, or social (Stalnaker, 2017, p. 903).

Rooted in this distinction, Speaks (2010) develops a framework that characterizes different approaches to meaning in terms of semantic and metasemantic theories. Among the most relevant semantic theories for the present paper, the author identifies the theory of reference, commonly summarized under the slogan *meaning is reference*. This theory offers a framework for assigning meaning to expressions based on their contribution to the truth value of a sentence. On the metasemantic side, Speaks discusses the causal origin theory, initially developed by Kripke (1980) in *Naming and Necessity*. Kripke argues that the meaning of a proper name is explained by its

historical usage, which includes both (a) an initial "baptism"—the moment when the name is first introduced to refer to a specific entity—and (b) the successful transmission of the name’s usage from one speaker to another. Putnam’s externalism arises in the context of debates concerning Kripke’s causal origin theory. Meanwhile, Putnam’s externalist framework asserts that the facts justifying the meaning of natural kind terms are sociological, rooted in the practices of the linguistic community. The key point, as Speaks observes, is that semantic theories address semantic facts. In contrast, metasemantic theories involve various types of facts—sociological, historical, or psychological—depending on the particular focus of the theory.

3.2.2 Semantic and metasemantic methodologies

In this section, I argue that in the literature, the question of whether LLMs generate meaningful outcomes, that is, The Problem, has been tackled by using both semantic and metasemantic methodologies. This distinction does not depend directly on whether the approach to The Problem adopts a semantic or metasemantic theory of meaning. As I will demonstrate later in this chapter, Mandelkern and Linzen (2024) adopt a metasemantic theory of meaning, Putnam’s Externalism, yet still employ a semantic methodology. Instead, this distinction depends on how the theory of meaning is applied. To clarify this point, I first revisit the three-step framework employed in nearly all attempts to solve The Problem, which was reconstrued in the former chapter.

In this three-step framework, the first step involves adopting a theory of meaning in the form of Thesis (1). The second step consists of determining, from Thesis (1), Condition (2), which explains why a given expression means whatever it means under Thesis (1). Following the distinction between semantic and metasemantic theories introduced in the former section, it can be said that Condition (2) constitutes a metasemantic fact. The final step involves applying Condition (2) to LLM outcomes. The distinction between semantic and metasemantic approaches to The Problem depends on whether condition (2) is considered beyond the scope of thesis (1) or not. On the one hand, semantic attempts to solve The Problem interpret thesis (1) as a semantic principle. Thus, condition (2), as a metasemantic fact, is outside the scope of thesis (1). Accordingly, semantic attempts to solve The Problem often need to provide a novel extension of the foundational theory of meaning underpinning thesis (1). This novel extension, condition (2), consists of a metasemantic novel addition to the foundational theory of meaning adopted.

On the other hand, metasemantic methodologies interpret thesis (1) as a metasemantic principle. In this way, condition (2) is not outside the scope of thesis (1). Works that follow this strategy in general are faithful to the foundational work adopted, and they usually do not propose an extension to encompass condition (2). Instead, condition (2) is just taken from the theory of meaning adopted (see, for example, Grindrod, 2024).

The fundamental difference between these two approaches, semantic and metasemantic, lies in their respective focal points. While semantic methodologies focus on determining the meaning of LLMs’ outcomes, metasemantic methodologies center attention on studying the facts in virtue of which such meaning is ascribed. Suppose again that for any given expression ϕ , $HA(\phi)$ represents instances of ϕ as spoken, written, or thought by a human agent, and $AA(\phi)$ denotes those generated by an artificial agent, such as LLMs. In the semantic approach to The Problem, the focal point is on the attribution of meaning to $AA(\phi)$. The central question is whether and to

what extent $AA(\phi)$ bears meaning. The meaning of $HA(\phi)$ serves as the starting point for this analysis. Those who adopt a semantic methodology, then, are concerned about whether $AA(\phi)$ means whatever $HA(\phi)$ means. Accordingly, semantic approaches treat Thesis (1) as a semantic principle that ascribes meaning to $HA(\phi)$.

More specifically, the strategy typically followed to answer this question involves adopting a principle that ascribes meaning to $HA(\phi)$, and investigating whether this principle can also be applied to $AA(\phi)$. By its very nature, this investigation inevitably involves analyzing the metasemantic facts that explain why the principle successfully ascribes meaning to $HA(\phi)$, and whether those same facts can justify applying the principle to $AA(\phi)$. However, because the adopted principles only ascribe meaning to $HA(\phi)$, it constitutes a semantic fact that does not directly address the metasemantic grounds underlying meaning. Authors following this strategy often address this lack by providing a novel extension of the foundational work adopted in the form of Condition (2). Thus, Condition (2) constitutes a metasemantic novel extension of the theory of meaning adopted.

On the other hand, metasemantic methodologies begin by examining the facts that ground the ascription of meaning to $HA(\phi)$. The central question, then, is whether the same facts that justify the attribution of meaning $HA(\phi)$ also account for the attribution or denial of meaning to $AA(\phi)$. In this case, the focus is on the explanatory power of a given metasemantic theory—originally developed to explain the facts that justify why $HA(\phi)$ means what it means—to conceptualize a new phenomenon, namely $HA(\phi)$. The strategy involves investigating whether the same underlying explanatory framework developed for conceptualizing $HA(\phi)$ can be applied to this new phenomenon $AA(\phi)$.

The approach typically followed involves two steps: (1) adopting a metasemantic theory that explains why $HA(\phi)$ means what it means, and (2) analyzing whether those facts could also serve as a justificatory framework for explaining why $AA(\phi)$ means what it means. The nature of the relevant metasemantic facts depends on the particular theory in question. Some approaches emphasize sociological factors, while others focus on historical or psychological ones. Fundamentally, these metasemantic proposals do not need to provide a novel extension to determine Condition (2), as it comes with the theory of meaning adopted. In consequence, these approaches tend to be more faithful to the foundational theory of meaning adopted (a good example of this methodology can be found in Grindrod, 2024).

3.2.3 Two Semantic attempts to address the Problem of Meaning in LLMs

In this section, I illustrate two attempts to address The Problem by adopting semantic methodologies. As noted earlier, whether the employed method is classified as semantic or metasemantic does not depend on the theory itself, but rather on whether Condition (2) is considered as falling outside or within the scope of Thesis (1). To illustrate semantic methodologies, I offer a novel, systematic reconstruction of the argument presented in Bender and Koller (2020) and Mandelkern and Linzen (2024). The first example employs a semantic methodology, incorporating a semantic theory of meaning; the second example, on the other hand, adopts a metasemantic theory of meaning. In both cases, Condition (2) is considered to fall outside the scope of Thesis 1. This is illustrated by the fact that both proposals require extending the original theory of meaning

adopted in a novel way to define Condition (2).

Although this thesis does not provide a strong argument in favor of a metasemantics methodology, it does demonstrate that semantic methods in the two examples come with several problems. In the first case, the fundamental problem lies in expecting that a semantic theory of meaning would explain metasemantic facts that are beyond the explanatory scope of the former. In the second case, the fundamental problem lies in not being faithful to the original foundational work adopted, Putnam’s Externalism, by providing a semantic interpretation of what it was originally proposed as a metasemantic theory of meaning.

Meaning is not reference in LLMs

In their influential paper *Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data*, Bender and Koller (2020) starts by distinguishing between form and meaning. Form refers to any observable realization of language, and meaning denotes the correlation between form and aspects external to the language. The central thesis is that a model trained on form cannot learn meaning (Bender & Koller, 2020, p. 5187). To illustrate this point, the authors propose a thought experiment called the octopus test. This thought experiment involves two characters on separate islands communicating via telegrams until an octopus intercepts their messages. Over time, the octopus observes the exchanges, identifies patterns, and recognizes contextual similarities between certain words. Eventually, the octopus begins interfering in the conversation, generating responses and pretending to be one of the characters.

Even though the octopus repeats the pattern of the conversation and generates answers that could seem like those of the other characters, in a critical situation where one of the characters is in trouble and asks for a way to save their life from an angry bear, the octopus won’t be able to offer a suitable answer to save the other person life. Such a task requires the ability to accurately map between words and real-world entities, a capacity the animal lacks. Even though the octopus can understand that “coconut” and “rope” are semantically similar, it cannot grasp their meaning since it does not have access to the referents of these objects. LLMs, such as the octopus, lack the availability of having epistemic access to the reference of our words. Bender and Koller conclude by claiming LLMs cannot understand meaning, even though they can represent semantic similarities and general patterns present in the data (Bender & Koller, 2020, p. 5189).

Following the three-step strategy formulated in the former chapter, Bender and Koller’s argument can be schematically reconstructed as follows:

1. *Thesis (1)*: The meaning of a symbol is given by the referent of that symbol in the world.
2. *Condition (2)*: Understanding meaning requires accurately mapping between words and real-world entities.
3. *Application of Condition (2) to LLMs*: Since LLMs are trained solely on form, they cannot satisfy the epistemic condition (2).

In this reconstruction, it can be seen how the semantic theory of meaning is adopted, the theory of reference, grounds Thesis (1). To apply this theory of meaning to The Problem, the authors move beyond Thesis (1). The key point is that Thesis (1), by itself, does not indicate whether it serves as a semantic principle for LLMs’ outcomes. The authors, thus, require going beyond this thesis and defining an epistemological condition for semantic meaning. Condition (2),

hence, represents an extension of the theory of meaning adopted, which serves as the standpoint from which to evaluate whether LLMs' outcomes are meaningful. Fundamentally, this condition is no longer a semantic fact, but rather a metasemantic justification of Thesis (1). In other words, it explains the epistemological ground in virtue of which Thesis (1) works for human speakers. Then, in the next step, the authors demonstrate that this epistemic ground cannot be extended to justify the application of Thesis (1) to LLMs' outcomes, thereby providing a negative answer to The Problem.

Based on this reconstruction of Bender and Koller's argument, it becomes clear how this proposal adopts a semantic methodology grounded in the theory of reference, a semantic theory of meaning. Piantadosi and Hill (2022) elaborates a critical stance toward this argument by claiming that the semantic principle *meaning is reference* (Thesis 1) fails to account for a substantial class of expressions that cannot be explained in terms of their referents. In this section, I suggest that a deeper methodological flaw underlies the argument. Bender and Koller attempt to address The Problem by using a semantic fact (Thesis (1)) as a normative principle to determine metasemantic facts, namely, Condition 2 and Point 3. Semantic theories of meaning are theoretical constructions that result from idealizing a portion of a language to explain the meaning of a subset of expressions of that language. Metasemantic theories attempt to explain semantic principles in non-semantic terms by reducing them to psychological, mental, historical, and other types of facts. Thus, the flaw in this argument lies in adopting a semantic theory of meaning as a framework to determine metasemantic facts.

To highlight the issue I recognize behind adopting the semantic theory of meaning as a framework determining metasemantic facts, I suggest extending Speaks (2010)'s analogy between semantic and metasemantic investigations, and the task of a social anthropologist. More specifically, to develop the distinction between semantic and metasemantic theory of meaning, Speaks (2010) develops an analogy between these two sorts of theories and the task of a social anthropologist. Suppose that a social anthropologist is studying the table manners of a distant community. Two different sorts of investigations are at play. One task in which the anthropologist might be interested is describing these table manners by developing a system of categories and identifying which specific manners fall into which category. Similarly, constructing a semantic theory of meaning entails defining foundational principles for ascribing meaning and explaining how different expressions derive their meanings based on these principles.

In contrast, the anthropologist might also be interested in understanding the nature of table manners. This investigation would require examining how the rules governing table manners come to hold authority within the community. In this context, the key question is: *What fact about the community explains that people adhere to this particular system of table manners rather than another?* Here, the focus shifts from the properties of table manners to the underlying actions or norms that justify those properties. This type of inquiry parallels metasemantic investigation, which explores the underlying facts that justify why a principle for ascribing meaning is adopted over another. This analogy is proposed by Speaks (2010) to develop the distinction between semantic and metasemantic theories of meaning.

Moving towards Bender and Koller's argument, I suggest the following extension of Speaks (2010) analogy. Suppose that the social anthropologist is examining a fundamental *Eating Slowly*

rule based on social status, where the highest-status guest is expected to finish eating first, followed by guests of lower status. As a consequence of respecting these rules, the highest-status guest always finishes first among the other guests. The anthropologist might then ask: In virtue of what fact about the community has this particular *Eating Slowly* rule, based on social status, come to govern their table manners? To address this question, the anthropologist might hypothesize that the rule itself determines the social facts that justify its governance over the community. For instance, they could observe the consequences of this rule within the community and notice that higher-status guests often skip dessert to respect the rule and finish first. Based on this reasoning, the anthropologist might conclude that the social fact explaining why the Eating Slowly rule has come to govern their manners is that higher-status individuals skip dessert at social gatherings.

This analysis illustrates a confusion between two distinct types of inquiry. The facts that explain why the *Eating Slowly* rule governs this community should be traced to factors like the society's hierarchical structure and how that structure influences table manners. Investigating the *consequences* of applying the rule itself within the community does not uncover the *grounds* for why the rule exists. The confusion arises from assuming that the *Eating Slowly* rule determines the facts that justify its status as a governing social norm. However, studying the properties of the *Eating Slowly* rule and how it might affect the community, on the one hand, and investigating the facts that justify its adoption as a social norm, on the other hand, are fundamentally different types of inquiry.

A similar conceptual confusion arises when addressing The Problem by adopting a semantic theory of meaning as a normative framework. Just as our anthropologist might expect the social consequences of the Eating Slowly rule to explain why the rule exists, Bender and Koller, in this reconstruction of their argument, take Thesis 1 as determining the facts that belong to a different level of analysis, Condition 2, and, in consequence, Point 3. As it is beyond the explanatory scope of the social consequence of this rule, the grounds in virtue of which the rule exists, it is beyond the explanatory scope of a semantic theory of meaning to determine metasemantic facts that ground such a thesis, such as Condition 2.

From a more general point of view, this conceptual confusion can be explained as follows: while investigating principles for ascribing meaning and other fundamental semantic notions, the domain of inquiry is a given language or a portion of it. Semantic theories are theoretical constructions that result from idealizing a portion of a language. On the other hand, metasemantic investigations aim to justify semantic properties by reducing them to non-semantic facts. Thus, the domain of the inquiry is no longer a portion of language but a class of social, psychological, or historical facts, among other options. The flaw in this semantic attempt to solve the grounding problem lies in assuming that a theoretical construction initially derived from idealizing a portion of a language can define facts that belong to a metasemantic level of inquiry. This strategy assumes that these theoretical constructions can determine those non-semantic facts that answer whether LLMs generate meaningful outcomes.

LLMs and meanings in the linguistic community

Mandelkern and Linzen (2024) presents the first attempt to address The Problem relying on Putnam's theory of meaning. The authors start their argument with a thought experiment:

suppose a group of ants winding through the sand form a pattern resembling the sentence *Peano did not prove that arithmetic is incomplete*. While reading the pattern, Lucy receives a message from her university friend stating *Peano did not prove that arithmetic is incomplete*. This raises the paper’s central question: when an LLM produces the word “Peano”, is it genuinely referring to Peano, like Lucy’s friend does in their message, or is its use more like the incidental Peano pattern formed by the ants? Inspired by Putnam (1975), the authors propose that the meaning of “Peano” is fixed not by Lucy’s individual beliefs concerning the famous mathematician but rather by “standing in the right kind of natural history to the referents” (Mandelkern & Linzen, 2024, p.1192).

More specifically, following Putnam’s theory of meaning, they argue that, in the case of historical names, meaning originates in baptism-like events in which a historical figure is given a name. This initiates a historical chain of events connecting the original naming event to the current usage of the name. In other words, the speech community fixes the meaning of “Peano” through this initial baptism event, and a historical chain links Lucy’s friend’s use of the word to that event. Within this framework, the pertinent question is no longer whether an LLM has the functionality to ground an expression in the correct reference in the world. Instead, what matters is whether the LLM’s use of “Peano” traces back to Peano through the right kind of causal historical chains required for reference (Mandelkern & Linzen, 2024, p. 1196). This perspective reframes The Problem by emphasizing the role of training data and the possibility that it carries fundamental information about these historical chains. The authors further argue that The Problem, as formulated by Bender and Koller, relies on a naïve conception of data as mere strings of symbols. However, they contend that the data used to train these models is far more complex and might be suffused with a history of meaningful use. In this sense, while the authors do not offer a definitive answer to The Problem, they suggest that the LLM’s use of the word “Peano” is closer to Lucy’s friend’s use than to the pattern created by the ants in the sand.

Following the three-step strategy reconstructed in the former Chapter, this proposal can be schematically summarized as follows:

1. *Thesis (1)*: Meaning is fixed by the linguistic community.
2. *Condition (2)*: Producing meaningful expressions requires “standing in the right kind of natural history to the referents” (Mandelkern & Linzen, 2024, p. p.1192).
3. *Application of Condition (2) to LLMs*: Given the data in which these models are trained, LLMs stand in the right kind of natural history to the referents.

As noted, two characteristics help identify when an attempt to solve The Problem lies within a semantic methodology. One of these characteristics is defining Condition (2) as a novel metasemantic extension of Thesis (1). The second characteristic is interpreting Thesis (1) as a semantic theory of meaning. In what follows, I argue that Mandelkern and Linzen (2024), despite adopting a metasemantic theory of meaning, still depends on a semantic methodology.

In the first place, although Thesis (1) represents Putnam’s externalism, condition (2) does not constitute a faithful adoption of Putnam’s externalism. The authors claim that LLMs’ outcomes are meaningful if LLMs use of “Peano” traces back to Peano through the right kind of causal historical chains required for reference. In other words, LLMs’ capability to produce meaningful output depends on the ability to trace an expression back to its baptism-like event via the right

kind of causal-historical chains. This application presupposes condition (2), under which speakers produce meaningful expressions as a consequence of standing in the right kind of natural history to the referents. This movement would only be a faithful adoption of Putnam’s externalism if a similar condition holds for the speaker. That is, speakers produce meaningful output given the ability to trace an expression back to its baptism-like event via the right kind of causal-historical chains. Although other forms of externalism ascribe to the speaker some sort of capabilities or psychological features as a fundamental element for fixing reference, this is not the case for Putnam’s externalism. In support of this reading is the fact that, at the end of the paper, the authors abandoned Putnam’s externalism and adopted a broader perspective on externalist theories of meaning, in which the speaker’s intention to refer starts to play an explanatory role again. Specifically, they close the paper claiming that the discussion regarding The Problem should move towards analyzing whether LLMs can belong to the linguistic community by having some condition with a lower theoretical burden regarding the intention to refer to.

Why do the authors fail to be faithful to the externalist version as developed by Putnam while addressing The Problem? Why do they move the discussion closer to the mentalist spectrum of the discussion, even though they adopted a theory that has the theoretical advantage of not depending on mentalist features while explaining meaning? A possible reason is that they use Putnam’s externalism as a semantic theory of meaning that works as a semantic theory defining Thesis (1) and other versions of externalism as a metasemantic theory that defines condition (2). In other words, the difficulty lies in that they start their investigation at a semantic level, asking for the meaning $AA(\phi)$ rather than the fact that justifies such meaning. Thus, they begin by adopting a first thesis that will serve as a semantic thesis, and then analyze the grounds on which this thesis stands.

The fundamental problem is that, given that they answer Bender and Koller’s semantic formulation of the problem, they inherited a semantic methodology, despite adopting a metasemantic theory of meaning. In other words, while following Bender and Koller (2020), they use Putnam’s externalist theory to explain the semantic meaning of $AA(\phi)$ rather than on an investigation of the metasemantic facts that explain the meaning of $AA(\phi)$, for it has been intended initially to explain. More specifically, their primary focus remains on ascribing meaning to the $AA\phi$, rather than investigating the facts that explain its meaning. More specifically, the authors ask whether an LLM’s use of an expression can be traced back to the baptism-like event for that expression through the appropriate causal-historical chains. They argue that LLM outputs can be considered meaningful if this condition is met. In other words, the ability to trace an expression back to its baptism-like event via the right kind of causal-historical chains serves as a condition for ascribing meaning to LLM outputs rather than as an explanation of the facts that justify such meaning. In this sense, Putnam’s externalism seems to be used more in terms of a semantic theory that defines a condition for ascribing meaning to a kind of linguistic token, namely $AA\phi$, rather than an explanation of the fact that justifies the meaning of $AA(\phi)$.

3.3 Ambitious and Modest Scope of Application of Putnam's Theory of Meaning

In this section, I introduce the distinction between an ambitious and modest scope of application of Putnam's Theory of meaning in three fundamental steps. First, I present this externalist picture of meaning as developed by Putnam (1975) in the well-known Twin Earth Thought Experiment. Second, I examine two different interpretations of this famous thought experiment: a radical, metaphysical interpretation and a moderate, non-metaphysical interpretation. Finally, I show that under the second interpretation, explicitly supported by Putnam in *Realism with a Human Face* (1992), *Words and Life* (1994), the classic problem of how language "hooks" onto the world is no longer a valid challenge. In particular, in his latter writing on the topic, Putnam argues that this challenge only arises if we overlook the role of the division of labor in fixing meanings within the linguistic community. Based on this second interpretation of Putnam's theory of meaning, I then define what constitutes a moderate scope of application and contrast it with an ambitious one.

3.3.1 Twin Earth thought experiment

The classical theory of meaning, such as reconstructed by Putnam, rests on two fundamental assumptions. First, knowing the meaning of a word is being in a certain internal psychological state. Second, intension determines extension, which means that the concept associated with a term determines what elements fall under it. Putnam aims to show that these two assumptions are not jointly satisfied by any notion of meaning, thereby demonstrating that the classical theory of meaning rests on a false premise (Putnam, 1975, pp. 134-36). The author's main thesis is that two speakers can be in the same psychological state concerning a word w even though the extension of the word w in the idiolect of one speaker is different from the extension of w in the other speaker's idiolect. Consequently, intension, defined as a psychological state, does not determine extensions.

To show his point, Putnam introduces the well-known Twin Earth thought experiment. Suppose there is a Twin Earth that is exactly like the Earth but with only one difference: the inner chemical structure of what the inhabitants of Twin Earth call water is not H_2O , but instead XYZ. Two important remarks to understand this thought experiment are that (i) XYZ in normal conditions is indistinguishable from water, and (ii) in Twin Earth, they also use the word *water* to talk about XYZ. Now, consider the year 1810, a time when Chemistry was not yet fully developed. Suppose Oscar₁ is an inhabitant of Earth, and Oscar₂ is his *doppelgänger* living on Twin Earth. The key point of the thought experiment is that even if Oscar₁ and Oscar₂ have the exact same beliefs concerning water, they understand the term in different ways since what Twin Earth inhabitants call water is not H_2O , but instead XYZ. Intension, defined as a psychological state, thus, does not determine extension.

Putnam goes further by positing the hypothesis that extensions are determined by the sociolinguistic community to which the speaker belongs. For instance, the extension of the word *gold* is not determined by the psychological states of individual speakers regarding the concept of gold. Instead, speakers are part of a linguistic community that operates under a particular

division of labor that distinguishes between expert and non-expert. This division designates which group within the community is responsible for determining the extension of *gold*.

On the one hand, extensions are thus established by the linguistic community, with the division of labor playing a central role in determining who is tasked with discovering the true nature of natural kinds or providing accurate definitions for certain artifacts. On the other hand, even though the concept of intension, as developed in the traditional philosophy of language, no longer applies, there is still room for studying individual competence within this framework. The traditional problem of meaning is still divided into two distinct dimensions: one that studies how extensions are fixed, and another that examines individual competence.

If this is correct, and I think it is, then the traditional problem of meaning splits into two problems. The first problem is accounting for the determination of extension. Since, in many cases, extension is determined socially and not individually, owing to the division of linguistic labor [...] Solving it would involve spelling out in detail exactly how the division of linguistic labor works. [...] The other problem is to describe individual competence. Extension may be determined socially, in many cases, but we don't assign the standard extension to the tokens of a word *W* uttered by Jones no matter how Jones uses *W*. Jones has to have some particular ideas and skills in connection with *W* in order to play his part in the linguistic division of labor. Once we give up the idea that individual competence has to be so strong as to actually determine extension, we can begin to study it in a fresh frame of mind. Putnam, 1975.

The Problem encompasses two distinct dimensions of analysis: extensions and individual capability. Investigating the meaning of a word *w* involves studying the division of labor within the community to identify which group is responsible for establishing the meaning of *w*. The second dimension concerns individual competence, which refers to the linguistic knowledge that the community expects a speaker *S* to possess regarding an expression *w*. This linguistic knowledge partially depends on the role that the speaker occupies within the linguistic community.

3.3.2 Two interpretations of the Twin Earth thought experiment

Following Rudd (1997) on *Two Types of Externalism*, there are two primary interpretations of Putnam's Twin Earth argument, each of which explains, following different paths, the thesis that the linguistic community fixes the meaning of our terms. These interpretations start by offering different perspectives on why the liquid that the inhabitants of Twin Earth call "water" is not water but something entirely different. The first interpretation adopts a more radical, metaphysical approach. According to this view, Putnam's thought experiment shows that natural kinds have intrinsic, objective structures, and it is the role of science to uncover these true structures. Consequently, classification established by science should reflect the actual structure of the world. Following this view, scientists have discovered that the correct structure of water is H₂O, thereby fixing the meaning of the word "water". Thus, even though the inhabitants of Twin Earth use the same word for a substance that superficially resembles water, they hold a mistaken belief that it is not H₂O but XYZ.

A second moderate and non-metaphysical interpretation situates Putnam's thesis concerning how meaning is fixed as a descriptive claim about the prominent role of scientific practices in determining how we classify the world. Within this interpretation, XYZ is not water since we accept the validity of the scientific practices that fixed the meaning of water as H₂O (Rudd, 1997, p. 504). To prevent Putnam's thesis from being reduced to a metaphysical interpretation, one should avoid the notion that scientific practices fix meaning because they are responsible for discovering the inner structure of nature. On the contrary, in this interpretation, the thesis that science fixes the meaning of our words is merely a descriptive claim about how the division of labor operates within our linguistic communities. This second interpretation has a fundamental characteristic: it does not aim to solve the challenge of how language hooks into the world. On the contrary, this picture of meaning "[...] presupposes language users who are already in contact with a world, not minds in the abstract or languages in the abstract which somehow have to be related to a world, which have to, as we say, find a way to hook on to a world" (Putnam, 1994, p.78-9).

To summarize these two interpretations, in the metaphysical interpretation of Putnam's externalism, water has an inner structure, and scientific practices—specifically chemistry—are responsible for discovering it. In doing so, they fix the meaning of our term "water". In contrast, under the non-metaphysical interpretation explicitly supported by Putnam (1983, 1992), given the actual division of labor, chemistry is one branch of science responsible for fixing the meaning of the natural kind "water".

In this thesis, I adopt the second non-metaphysical interpretation of Putnam's Externalism to address The Problem. In contrast with an idealized or abstract picture of meaning, which begins by looking for mechanisms by which language "hook on to a world", this non-idealized interpretation starts by presupposing language users and the meanings already in use within the linguistic community. In other words, it takes for granted ascriptions of meaning as they occur in the actual linguistic practices. This raises an important question: What is the role of a theory of meaning if it does not aim to explain how language hooks onto the world? And, more importantly, what shape does The Problem take when conceptualized from this non-idealized view of meaning? In the next section, I answer these two questions.

3.3.3 Why Language does not need to hook onto the world

In *Why is a Philosopher?*, Putnam focuses on the central philosophical problem in the analytic tradition of how language hooks onto the world. Rather than taking a side in this extensive debate, the author examines the very ground upon which this problem is formulated. Putnam argues in this writing that this challenge arises from a misguided conception of language, in particular, one that fails to account for two fundamental aspects of meaning: the division of linguistic labor and the contribution of the environment in fixing reference (Putnam, 1992, p 110). The traditional conception that overlooks these two fundamental aspects has given room to an individualistic and aprioristic conception of meaning. First, an individualistic picture of meaning overlooks a fundamental phenomenon that explains meaning, namely, semantic difference. Speakers semantically defer to other speakers while fixing the meaning of their own terms. Secondly, the traditional picture of meaning presupposes that our concepts contain analytic

truth. Putnam concisely expressed this idea in the following terms.

The seventeenth-century thought of concepts as entities immediately available to the mind, on the one hand, and capable of fixing reference to the world, on the other hand. In this picture, the concept «gold», for example, is in the mind of any speaker (even if he uses a Greek word, or a Latin word, or a Persian word) who can refer to gold. The extension, or reference, of the word “gold”, or “chrysos”, or whatever, is determined by the concept. This picture of language is both individualistic (each speaker has the mechanism of every word he uses in his own head) and aprioristic (there are analytic truths about the natural kinds we refer to, and these are contained in our concept). It is not hard to see that this picture does violence to the facts of language use and conceptual thought, however. Few speakers today can be certain that an object is gold without taking the object to a jeweler or other expert. The reference of our words is often determined by other members of the linguistic community to whom we are willing to defer. There is a linguistic division of labor which the traditional picture entirely ignores. Words acquire a direct connection with their references, not by being attached to them with metaphysical glue but by being used to name them. (Putnam, 1992, p. 108-9)

This picture of meaning, thus, involves two distinct dimensions. First, extensions: this dimension concerns questions such as why the linguistic community attributes a particular meaning to a specific word. Second, there are questions concerning the speaker's individual capabilities: why the community interprets a particular expression, uttered by a speaker, in a certain way. Both dimensions can be addressed by examining the division of linguistic labor. In particular, by examining the expert-non-expert dichotomy between spectra. To study the first dimension, one must take a given word and then identify the group within the community responsible for fixing the meaning of that word. To study the second one, one must take a word as uttered by a speaker, and investigate the speaker's role within the community to understand why their expression is attributed a particular meaning.

What is the scope of application of this picture of meaning? First, for a word w , this picture explains the extension of that word by studying which group in the linguistic community is in charge of fixing its meaning. Second, given a speaker S_0 , this picture explains the expected knowledge that other speaker S_1 will attribute to it depending on S_0 's role within the community. In particular, if S_0 occupies the role of an expert regarding a term T , and S_1 is a non-expert regarding T , then S_1 defers to S_0 regarding T . This picture does not solve the challenge of how language hooks into the world. It instead shows that this challenge is misguided once we take into account how labor distribution works regarding the task of fixing meaning in the linguistic community.

This scope of application contrasts with the Mandelkern and Linzen (2024) one. In particular, Mandelkern and Linzen (2024) applies the externalist picture of meaning to solve the question of how LLMs' outcomes hook into the world. This application presupposes that Putnam's theory of meanings provides a framework to explain how our words are grounded in the world. More specifically, if one treats the challenge of how language hooks into the world as a valid problem

and sees Putnam’s theory of meaning as offering a solution to it, then it makes sense to expect this same framework to address the analogous problem for LLMs, namely, how their output hook into the world. By contrast, if one interprets Putnam’s theory as demonstrating why the original challenge is misguided, as it arises only when we ignore the role of the linguistic division of labor, then it follows that the analogous problem for LLMs’ outcomes should also be regarded as misguided. From this perspective, the theory does not solve this challenge but dissolves it. The former proposal is what I call an ambitiously intended scope of applications, while the latter constitutes a modestly intended scope of applications of Putnam’s Theory of meaning.

3.4 A Metasemantic and Modest Application of Putnam’s Theory of Meaning

As we have seen, Putnam claims that the linguistic community fixes the meaning of our terms. Mandelkern and Linzen interpret this slogan as a principle that determines whether LLMs generate meaningful outcomes. Their interpretation constitutes both a semantic and ambitious scope of application of Putnam’s theory to solve The Problem. These two features, together, make explicit how their solution to The Problem is grounded in a *soft* cognitive notion of AI agency.

First, Mandelkern and Linzen begin with the general question of whether LLMs’ outcomes mean something (in alignment with their semantic methodology). Then, they conceptually refine this question in terms of how LLMs’ outcomes hook into the world (in alignment with their ambitious scope of application of Putnam’s Externalism). Finally, they answer this question by investigating whether these models can simulate properly the inner process by which a speaker enunciates a word in accordance with the history of use of that word in the linguistic community. If so, then these outcomes successfully hook into the world and, consequently, they mean something. Hence, their proposal manifests a light but yet cognitive feature: the expectation that LLMs, in order to generate meaningful outcomes, must simulate human modes of linguistic engagement (in this case the speaker’s capacity to generate a word in accordance with its history of use in the linguistic community) with the environment (in this case, the linguistic community). As I showed in the reconstruction elaborated in Chapter 1, this way of framing The Problem presupposes a soft cognitive notion of AI agency.

In contrast with Mandelkern and Linzen’s proposal, in this section, I elaborate on the metasemantic and modest application of Putnam Externalism. In doing so, I ground my interpretation with a non-cognitive notion of agency. More specifically, I first advocate for adopting Putnam Externalism based on a metasemantic methodology. This means that rather than being concerned about the semantic meaning of $AA(\phi)$, I am concerned about what sort of facts explain the meaning we ascribe to $AA(\phi)$. Thus, this investigation begins on a metasemantic level, and it takes for granted the meaning we ascribe to $AA(\phi)$. Second, Putnam Externalism does not explain how language must hook into the world, but rather shows why this challenge is dissolved once we take into account labor distribution in the linguistic community. In the interpretation of Putnam’s externalism that I adopt in this thesis, there is no matter of fact concerning whether an LLM’s use of the word “Peano” is closer to Lucy’s friend’s use or to an ant’s random pattern in the sand, as Mandelkern and Linzen, 2024 expects. However, there is a matter of fact concerning how LLMs

are being integrated into the labor distribution in the linguistic community. In other words, there is a matter of fact regarding the fact in the linguistic community in virtue of which we ascribe meaning to $AA(\phi)$. These two features, thus, lead me to ask *how* are LLMs (linguistically) engaging with the environment (in this case, the linguistic community)? This way of framing the problem, as remarked, presupposes a non-cognitive notion of AI agency (See Table 3.1).

But exactly what sort of facts about the labor distribution would explain the meaning we ascribe to $AA(\phi)$? Or, in other words, what sort of facts explain how LLMs are linguistically engaging with the linguistic community? Meaning involves two dimensions, extension and individual capabilities. In what follows, I provide an investigation into the types of facts that would explain how LLMs are being integrated into the linguistic community, based on these two different dimensions of analysis.

The next two sections are organized as follows. I start with a brief description of how the relevant dimension of meaning is defined in Putnam’s externalist picture. I then offer a short explanation of how I intend to apply this dimension to LLMs; finally, I apply it, highlighting the sort of facts that we must study to understand how LLMs are being integrated into linguistic communities. Although I do not provide a final answer to this question, I do suggest potential research lines.

It is worth noting that not providing a definitive answer to The Problem of Meaning does not imply failing to fulfill the objective of this thesis. The objective lies in creating the conceptual space to reconcile two seemingly opposing intuitions: namely, that LLMs can successfully engage in linguistic interaction, even though they do not simulate a cognitive process of meaning production. By framing the problem in terms of the question How are LLMs linguistically engaging with the environment (linguistic community)? I create such a space. In other words, by formulating the problem on the basis of both a non-mentalist theory of meaning and a non-cognitive notion of AI agency, I fulfill the objective of this thesis. Thus, answering this question in detail lies outside the scope of this work. Nonetheless, I conclude this section by suggesting potential lines of research that might emerge from pursuing this approach.

3.4.1 Extensions: the meaning of “meaning” in the age of LLMs

The case of the synthetic dogs

To clarify how I apply the dimension of meaning involving extensions to LLMs, I first introduce the case of the synthetic dogs investigated by Putnam in “Aristotle after Wittgenstein”. In this work, Putnam asks the reader to imagine a scenario where technology advances to the point that it becomes possible to create synthetic dogs, that is, dogs that have the same DNA as natural dogs but lack the evolutionary history since they are entirely produced in laboratories. This raises the question: are these synthetic dogs real dogs?

Putnam’s answer is fairly straightforward. From an evolutionary perspective, synthetic dogs are not considered real dogs because they lack the essential evolutionary history that, according to evolutionary biology, constitutes the criteria for the individuation of the natural kind *dog*. However, from a molecular biologist’s point of view, where individuation is based on DNA, synthetic dogs and natural dogs would be categorized as the same kind. The point of this example is that there is no absolute viewpoint, or final scientific field, that fixes the meaning of a word.

Instead, a more accurate picture shows that different branches of science fix the extension of the same word from different perspectives.

Putnam goes further and remarks that, despite the absence of an absolute viewpoint, not every conceptualization of a given word fixes its meaning. Only those that exhibit a sufficient degree of rationality in the conceptualization and are relevant for the linguistic community can do so. Putnam illustrates this point with the example of how to wash a dog. This task is neither relevant nor aligned with the standard of rationality present in the scientific community. While such knowledge might be important for a dog owner, it does not constitute either a relevant or a rational enough contribution to fix the meaning natural kind *dog* within the scientific community.

How synthetic dogs change our understanding of what it means to be a *dog*

Now, suppose one concludes that synthetic dogs are not real dogs because they lack the necessary biological history to be a dog. What, according to Putnam, is wrong with this conclusion? This example shows that there is no matter of fact concerning whether these technologies are or are not real dogs. Whatever answer we provide, it will be strongly tied to the theoretical background we adopted. Thus, one cannot conclude that synthetic dogs do not consist of real "dogs" within the area in which they have been designed, synthetic technologies, because, from an evolutionary perspective, they do not share the same DNA as real dogs. Similarly, we should not conclude that synthetic dogs do not constitute an advance in the area of synthetic technologies because they are not real dogs. Such a conclusion mistakenly applies the standards of the evolutionary biology perspective to assess the advances in a different area, synthetic technologies. In other words, this shift would imply extrapolating the standard under which a given scientific field fixes the meaning of the word "dogs" to assess a different area, which comes with other standards.

However, although there is no matter of fact concerning whether synthetic dogs are real dogs, I suggest that the externalism picture of meaning is not totally indifferent to this topic. In particular, I claim that the externalist picture reveals that there is a matter of fact concerning how the extension of the natural kind *dog* has changed since the introduction of synthetic dogs in the linguistic community. My argument is as follows: if one is interested in how the introduction of synthetic dogs might modify the understanding of the word "dogs" in the linguist community, a good starting point is the linguistic division of labor surrounding the word's extension. By analyzing how the division of labor has changed since the introduction of these synthetic dogs, one might observe how a new group of experts (researchers in synthetic technologies) is contributing to the task of fixing its meaning. Accordingly, this new group of experts is fixing the meaning of the word from a new perspective, modifying its extension and, in consequence, the understanding of the word within the linguistic community.

The final picture of meaning, thus, looks as follows: within the externalist theory of meaning, there is no matter of fact about whether dogs are real dogs. But there is a matter of fact about how the introduction of synthetic dogs changes the linguistic labor distribution around the word "dogs". Since their introduction, a new relevant group of experts within the scientific community, researchers in synthetic technology, is fixing the extension of the word "dogs" from a novel perspective. This means there is a matter of fact regarding how the introduction of these new technologies changes the extension of the word "dogs" in the linguistic community. In other

words, Putnam's theories of meaning provide a framework for investigating how these synthetic dogs transform our understanding of what it means to be a *dog* within the linguistic community.

From synthetic dogs to LLMs

Suppose now we ask: are LLMs real speakers? For the sake of the argument, suppose we rephrase this as: Do LLMs genuinely produce meaning? Based on the example of the synthetic dogs, there are two different routes for addressing this question. First, one might try to answer it by adopting theories that explain what constitutes being a producer of meaning from a theory external to the area in which LLMs have been designed, and, then, assessing to what extent LLMs resemble them.

As we have seen in the first chapter, the standard approach to The Problem follows this route. However, the example with the synthetic dogs shows that one should be cautious while taking this route. Although this route might bring several interesting insights, it is important to be aware of its scope. For instance, it might be misguided to determine that LLMs do not genuinely produce meaning based on whether they met a criterion external to the specific domain of expertise in which such technologies were designed. Similarly, it might be misguided to conclude that such technologies do not constitute a real advance in the field in which they have been developed because they cannot satisfy those external criteria. From the externalist picture of meaning, there is no matter of fact concerning whether LLMs genuinely produce meaning in an absolute sense, and whatever answer we provide to such a question, it would be tied to the theoretical background adopted.

However, here I do not aim to explore this path further, but rather propose a second route: conceptualizing how the introduction and increasing use of LLMs-based technologies might transform our understanding of what it involves *producing meaning* within the linguistic community. As remarked, within this externalist framework, although there is no matter of fact concerning whether LLMs are genuinely produce meaning, there is a matter of fact about how the introduction of LLMs changes the linguistic labor distribution around the word "production of meaning". In particular, it might be the case that, since their introduction, a new relevant group of experts within the scientific community is fixing the extension of the word "production of meaning" from a novel perspective. If this is the case, then, there is a matter of fact regarding how the introduction of these new technologies changes the extension of the word *producing meaning*, and, in consequence, our understanding of it within the linguistic community.

Thus, I propose asking how the labor distribution operating around this word has been changing since the introduction of LLMs into our linguistic communities. Answering this question would involve an exploration that is beyond the scope of this thesis; however, I would like to provide a possible hypothesis to explore it. In recent years, the word "meaning" has been increasingly used in a specific subfield of NLP, explainable AI (xAI) techniques for LLMs. xAI techniques were initially developed with the goal of explaining the behavior of opaque AI systems in human-understandable terms (Budding & Zednik, 2024; Zednik, 2021; Zhao et al., 2024). Deep learning systems are classified as opaque because of the challenges inherent in understanding their mechanisms. LLMs, likewise, are considered opaque systems that, in contrast with standard deep learning models, introduce an even higher degree of complexity to xAI techniques due to

their size, extensive training data, and vast number of parameters (Zhao et al., 2024)¹. The complexity of the internal mechanisms in these models makes it extremely challenging to explain their behavior in terms that are understandable to researchers, developers, and end-users as well. xAI techniques emerge intending to decrease LLMs' opacity.

Within this subfield, different evaluation paradigms have been systematically developed to interpret the behavior of these models in recent years for LLMs (see Linzen & Baroni, 2020; Mahowald et al., 2024; Marks & Tegmark, 2023; Wei et al., 2023; Wu et al., 2023; Zhao et al., 2024). Researchers working in this area often interpret xAI techniques as aiming to define or explain how LLMs produce meaning (see Budding & Zednik, 2024; Linzen & Baroni, 2020; Pavlick, 2023; Zhao et al., 2024). A potential route to explore, thus, is whether the introduction of LLMs gave rise to a new field that fixes the extension of the "meaning" or "production of meaning" from a new perspective. This would involve reconstructing the notion of "meaning" or "production of meaning" as it is being used into the sub-field of Natural Language Processing, xAI techniques for LLMs.

This question involves a shift in the perspective from which The Problem has been investigated in the philosophical literature. Rather than adopting, for example, a mentalist theory of meaning as a framework to determine whether LLMs' outcomes are meaningful or not, this question involves asking how the introduction of a new language technology, LLMs, might change how labor distribution is operating around this word and, consequently, modify the extension of the word itself within the linguistic community.

This new approach to The Problem can be placed at the very end of the non-cognitive side of the spectrum. In contrast with Mandelkern and Linzen (2024), I am not applying Putnam's theory of meaning to determine whether LLMs simulate the inner process by which speakers produce meaning. Instead, I suggest using this theory to investigate the specific modes of linguistic engagement of these technologies with the linguistic community. As this first dimension of meaning illustrates, one way to explore how LLMs are being integrated into the linguistic community is by examining how these new technologies might modify our understanding of what constitutes *meaning production*.

3.4.2 Individual Capabilities: the role of LLMs in the linguistic community

The phenomenon of semantic deference

So far, I have been focusing on the concepts of linguistic community, labor distribution, and extension. But Putnam's externalism introduces a second crucial dimension: individual capabilities. These refer to the linguistic knowledge that a community attributes to a specific speaker based on their role within that community. As Putnam remarks while introducing this concept: "extension may be determined socially, in many cases, but we don't assign the standard extension to the tokens of a word *W* uttered by Jones no matter how Jones uses *W*. Jones has to have some particular ideas and skills in connection with *W* in order to play his part in the linguistic division of labor" (emphasis added, Putnam, 1975, pp. 165-166).

The starting point for explaining the concept of individual capabilities is the claim that

¹For instance, as noticed, in the case of GPT-3, we are dealing with a model that has 175 billion parameters and uses 570 GB of text data for training (Wang et al., 2024).

speakers possess some form of linguistic knowledge. However, it no longer plays the traditional role of determining the extension of a word. On the contrary, the direction is reversed: it is the speaker's role within the linguistic community and their place in the division of linguistic labor that determines what kind of linguistic knowledge is expected of them. As Putnam observes, "few speakers today can be certain that an object is gold without taking the object to a jeweler or other expert. The reference of our words is often determined by other members of the linguistic community to whom we are willing to refer" (Putnam, 1992, p. 108).

This shift, from viewing a speaker's knowledge as the basis for fixing meaning to understanding their role in the community as what shapes the knowledge they are expected to have, is central to understanding *semantic deference*. Although Putnam does not introduce this notion under that name, it is clearly described in the passage cited above. This concept was further explored in more detail by Burge in his version of externalism for mental contents (Burge, 1979, 1986, 2007). Within this picture, thus, speakers play different roles in the linguistic community, and the linguistic knowledge we expect from a speaker depends on this role. Fundamentally, a speaker's role within the linguistic community determines whether other non-expert speakers will defer to them or not.

The linguistic community as a network of semantic deference

What can be concluded from the example of the jeweler? By using appropriately the term gold, the jeweler plays a linguistic role as someone responsible for distinguishing gold from non-gold materials. This role is justified because the jeweler uses the word "gold" appropriately. In this case, the community attributes linguistic knowledge concerning the extension of the word "gold" to the jeweler, but this attribution is restricted to a specific domain of expertise: jewelry. This domain restriction is crucial because there is no absolute expertise over the term gold; rather, expertise is relative to the particular perspective that underpins the field of science or area of knowledge to which the speakers belong. For instance, if a speaker is interested not in distinguishing gold from other materials but in determining its chemical structure, the relevant expert they rely on is no longer a jeweler but a chemist.

The linguistic community, hence, is organized around the labor distribution. This labor distribution distinguishes between experts and non-experts, although this distinction is relative to a specific field of knowledge. Non-expert Speakers rely on other expert speakers while determining the meaning of a word. Specifically, non-expert speakers semantically defer to experts to clarify the meaning of terms that fall within the scope of the latter's expertise. Semantic deference is justified when the relevant expert uses their "word" appropriately. I propose summarizing this picture of meaning as describing the linguistic community in terms of a *a network of semantic deference between expert and non-expert speakers*.

LLMs as a new sort of artificial speaker within a network of semantic deference

This picture illustrates that the speaker's role in the linguistic community plays a crucial role in determining the extension that other speakers assign to former's words. Fundamentally, when a non-expert recognizes an expert's authority, they semantically defer to the expert regarding concepts that belong to the latter's area of expertise. On this basis, I suggest an application of

the second dimension of Putnam's theory of meaning in terms of the following questions: What role, if any, do LLMs play in this network of semantic deference between speakers? In other words, how do users interpret the outcomes of LLMs, given the role these models play in labor distribution? Are LLMs closer to the role that experts play in the linguistic community? Or are they closer to non-experts? Fundamentally, if they play the role of experts, is this role justified?

These questions can be answered by reviewing whether LLMs are being applied to interact with other experts or with non-experts, and whether the topic in which they interact belongs to an expert or non-expert domain. For example, suppose a LLM-based chatbot is applied in expert domains (such as healthcare or finance) to interact with a non-expert (e.g., users, patients, or consumers). In that case, it is expected that the non-expert will semantically defer to the chatbot.

Although it is beyond the scope of this thesis to study how LLMs are being applied, I would like to suggest a possible hypothesis. A non-comprehensive first review of how LLMs are being applied in healthcare, an expert domain, shows that it might be possible to characterize the different application of these model into these two fundamental ways: cases in which they are applied to interact with another expert and potentially support expert opinion, and cases in which they are applied to interact with non-expert.

In the first place, LLM being applied to interact with other experts can be found in sub-areas such as diagnosis, support of health professionals, and research. An example of the first case is the use of LLMs to support and potentially improve the accuracy of expert diagnostics. For example, diagnosis Rau et al. (2023) developed an interactive chatbot that supports clinical decision-making and provides personalized imaging recommendations. Concerning the application of LLM for the support of health professionals, the literature proposes fundamentally that LLM could play a fundamental role in automatic repetitive administrative tasks, like medical reporting or summarizing medical interactions (p. 6 Haltaufderheide & Ranisch, 2024). Concerning applications of LLM in research, there are two fundamental uses: LLM application for research writing and LLMs for health research. Relating to the use of LLMs for writing, several authors remark on the potential of these models to automate literature review (Biswas, 2023; Cascella et al., 2023; Yang et al., 2023). For research purposes, Haltaufderheide and Ranisch (p 6 2024) also suggest that LLMs might potentially identify research targets, design experiments, facilitate knowledge sharing between collaborations, and support communication of study results. In all these cases, LLMs are being applied to interact with healthcare experts.

At the same time, this non-comprehensive review also shows that there are cases in which LLMs are being applied in healthcare to interact with non-experts. These uses can be fundamentally found in the sub-area of pre-diagnosis. For example, Table 1.1 constitutes a sample interaction of the recently introduced K Health agent. This product was released as the first intelligent AI knowledge agent in healthcare that uses patients' medical records for more personalized and accurate pre-diagnosis. It is designed to be the ideal first step in the diagnosis of general medical health care, in which the patient can solve fundamental doubts about what the most adequate next steps are given their symptoms (Bloch & Shaul, 2024). Similar attempts to develop an agent of this characteristic for pre-diagnosis can be found in the literature (Dave et al., 2023). Also, similar LLM-based systems for pre-diagnosis that interact with non-experts can be found

in mental healthcare (Blease & Torous, 2023). Also along the line of LLMs interacting with non-experts in healthcare, Yang et al. (2023) suggest that LLMs can be used to improve patients’ understanding and engagement with their medication.

Criterion	Cognitive End of the Spectrum	Center of the Spectrum	Non-cognitive End of the Spectrum
<i>(i) Notion of Agency</i>	A cognitive notion: AI agency requires simulating human-like intelligence.	A <i>soft</i> cognitive notion: AI agency requires simulating human-like linguistic interactions	A non-cognitive notion: AI agency replicates intelligent behavior by non-cognitive means.
<i>(ii) How to Investigate AI Agency</i>	ARI Thesis: Do AI technologies instantiate agency by successfully simulating intelligence?	Do AI technologies instantiate agency by successfully replicating humans’ modes of interaction with the environment?	ARA Thesis: How are AI technologies instantiating specific human modes of interaction with the environment by non-cognitive means?
<i>(iii) Theory of Meaning</i>	Mentalist theory of meaning	Putnam’s theory of meaning (Mandelkern and Linzen (2024))	Putnam’s theory of meaning (my proposal)
<i>(iv) Definition of Meaning</i>	A cognitive process	A specific mode of engagement with the environment (linguistic community)	A specific mode of engagement with the environment (linguistic community)
<i>(v) Conceptual Refinement of the Problem</i>	Do LLMs generate meaningful outcomes by successfully replicating the cognitive process underpinning meaning production?	Do LLMs successfully replicate human modes of linguistic engagement (grounding the use of a word in its history of use in the linguistic community) with the environment (linguistic community)?	How are LLMs linguistically engaging with the environment (linguistic community)?
<i>(vi) Key to Addressing the Problem of Meaning in LLMs</i>	Determining whether LLMs replicate the cognitive process of meaning production	Determining whether LLMs have specific modes of engagement with the linguistic environment by analyzing whether they ground the use of a word in its history of use in the linguistic community	Determining LLM’s specific modes of linguistic engagement by studying: (i) how they modify the labor division surrounding the word “meaning”, and (ii) which role, if any, they play in the network of semantic deference.

Table 3.1: Adaptation of Table 2.1. In the column *Non-cognitive End of the Spectrum*, I show the results of applying (a moderate and metasemantic version of) Putnam’s theory of meaning to solve the Problem; on the other hand, in the column *Center of the Spectrum* I show the results of results of applying (a ambitious and semantic version of) Putnam’s theory of meaning to solve the Problem

Suppose this non-compressive review is representative of what is going on in expert domains. That is, it is possible to distinguish between applications in which LLMs interact with experts and applications in which they interact with non-experts. What is the relevance of this distinction? From an externalist picture of meaning, the expected knowledge that other speakers attribute to an LLM-based language technology does not depend on whether they generate meaningful outcomes by simulating a cognitive process of meaning production. Instead, it depends on which role they are occupying in the linguistic division of labor in the community. This linguistic community is organized around a network of semantic deference in which the distinction between experts and non-experts is crucial. Thus, on the one hand, when these models are designed and applied to interact with non-experts, they are more likely to play the role of an expert. Hence, non-experts would defer to them while clarifying the meaning of a technical term. On the other hand, in cases in which these models are applied to support expert opinion in expert-expert interactions, it is less likely that the phenomenon of semantic deference occurs.

Fundamentally, if LLMs play the role of experts, such a role will only be justified if they use their terms appropriately. Otherwise, such a role would not be justified. The relevance of this conceptualization is that, if this picture is correct, two conclusions can be inferred. First, common failures in LLMs, such as hallucinations, factual errors, and similar issues, may have varying degrees of impact depending on the likelihood of semantic deference. This would imply that a proper understanding of how these language technologies must be applied requires careful consideration of the specific type of linguistic interaction in which LLMs are expected to participate. Second, semantic deference in LLMs might happen in the linguistic community even when it is not properly justified. In particular, these common failures, when occurring in interactions in which LLMs play the role of experts, would show that semantic reference is ill-justified.

The picture that results from applying this second dimension of meaning, individual capabilities, to LLMs, can be placed at the very end of the non-cognitive side of the spectrum. Rather than trying to determine whether LLMs simulate the inner process by which speakers produce meaning, I apply this theory to investigate the specific modes of linguistic engagement of these technologies with the linguistic community. Based on this second dimension of meaning, another way in which we can explore how LLMs are being integrated into the linguistic community is by studying which role within the expert-non-expert dichotomy they are playing within the network of semantic deference underpinning our linguistic communities. This investigation conceptualizes how the introduction of these new technologies is potentially transforming this network by introducing a new artificial speaker to whom non-expert speakers might be deferring.

Conclusion and Future Work

This thesis contributes to both the literature on AI agency and the problem of meaning in LLMs within the philosophy of AI. In particular, it elaborates a novel conceptual framework to reconcile two intuitions that, so far, have been presented as opposites in the literature on the problem of meaning in LLMs: the intuition that these technologies can successfully engage in linguistic interactions, on the one hand, and the intuition that they do not possess semantic knowledge nor they simulate the process by which humans produce meaning, on the other. Such a framework emerges from framing the problem by applying Putnam's Externalism. Fundamentally, this application is guided by a non-cognitive notion of AI agency.

Focusing on chapter 1, the main contribution lies in connecting the debate about agency and the problem of meaning in LLMs. In particular, I show how the latter problem can be understood as a specific instance within the broader discussion of agency. The route I followed to achieve this also allows for original sub-contributions: I provide the first systematic map of the literature on the problem of meaning in LLMs, and I also offer a novel characterization of the different attempts to solve it in terms of mentalist and non-mentalist approaches. On this basis, I make explicit the connection between these two debates by showing how mentalist approaches presuppose a cognitive notion of agency. In contrast, non-mentalist approaches presuppose a *soft* cognitive notion. I also demonstrated why, so far, no non-mentalist approach has framed the problem of meaning guided by a non-cognitive notion of AI agency, thereby setting the stage for the contribution of Chapter 2. Finally, I present an argument from parsimony, inherited from Floridi's framework for agency, in favor of a non-mentalist and non-cognitive framework for conceptualizing the problem of meaning in LLMs.

Regarding Chapter 2, the main contribution consists of addressing the gap recognized in Chapter 1: adopting a non-mentalist theory of meaning, Putnam's Externalism, and guiding its application to address the problem of meaning by relying on a non-cognitive notion of agency. As a result, I elaborated a novel approach to the problem of meaning that consists of investigating how these technologies are being integrated into our linguistic communities. In this context, there are also relevant sub-contributions worth mentioning: I distinguish between semantic and metasemantic methodologies for addressing the problem of meaning in LLMs, and I show how the tendency to adopt semantic methods can be traced back to one of the foundational papers on the topic, namely Bender et al. (2021). I also distinguish an ambitious application of Putnam's Externalism, focused on explaining how LLMs' outcomes hook into the world, and a moderate application, in which this theory only provides a framework for understanding how the division of linguistic labor within a community fixes meaning.

I argue that both features in conjunction, namely, a semantic methodology for addressing the

problem of meaning and an ambitious application of Putnam’s Externalism, provide the basis for grounding the problem of meaning in a *soft* cognitive notion of AI agency. From this perspective, solving the problem of meaning involves determining whether these technologies can simulate the internal processes by which speakers produce words in accordance with their history of use within the linguistic community. Accordingly, I make explicit how applying Putnam’s theory of meaning to frame the Problem, guided by a non-cognitive notion of AI agency, takes the form of a defense of a metasemantic, over a semantic methodology, and a modest, over an ambitious, application of Putnam’s Externalism.

In contrast with this semantic and ambitious application of Putnam’s Externalism, hence, I propose a novel approach grounded in the rejection of a cognitive notion of agency. In particular, rather than starting from the question of whether LLMs’ outputs have meaning, I begin with the question of in virtue of what facts we ascribe meaning to the outputs generated by these technologies. This represents a semantic methodology. Additionally, rather than expecting Putnam’s Externalism to explain how LLMs latch onto the world, I adopt this theory as offering a picture of how the labor of fixing meaning operates within a linguistic community, based on the distribution of linguistic labor. At most, then, this picture can help us understand the role LLMs are playing in the division of linguistic labor within the community. This is what I suggested to call a modest application of Putnam’s Externalism.

Additionally, I show how this modest application can still conceptualize the problem of meaning in LLMs in a novel way. In this framework, the task of fixing meaning can be understood from two perspectives: extension and individual capabilities. The first involves analyzing how different groups of experts within the linguistic community fix the meaning of our words. By applying this dimension to LLMs, I suggest exploring how various expert groups define the technical terms of meaning in LLMs within NLP and computational linguistics research. I propose that xAI techniques may play a fundamental role in fixing the meaning of *meaning* in LLMs. This leads us to ask how LLMs are being integrated into the linguistic community, particularly in terms of how these language technologies are introducing a new area of expertise that fixes the meaning of *meaning* from a new perspective. In other words, after the introduction of these technologies into the linguistic community, the extension of meaning has been modified, as it is now also being fixed in a technical context by different xAI techniques. Thus, one way in which we can explore how LLMs are being integrated into the linguistic community is by studying how these new technologies are modifying our understanding of the concept of *meaning* or, more precisely, *meaning production*.

The second dimension involves the notion of semantic deference and the expected individual capabilities of speakers based on their role in the linguistic community. In particular, semantic deference depends on the role that speakers play within the linguistic community. Fundamentally, when a speaker assumes the role of an expert in an interaction, non-experts tend to semantically defer to them. The question, then, is whether LLMs interact with users in the role of experts or non-experts. Although I do not answer this question, I suggest that it might be possible to consider cases in domains such as healthcare, where LLM-based chatbots designed for pre-diagnosis seem to play the role of experts. Specifically, in these contexts, patients are expected to defer to the chatbot regarding pre-diagnosis. If this holds, then by applying this second dimension to LLMs,

we can see how language technologies are being integrated into the linguistic community by adding a new agent to the network of semantic deference. Non-expert speakers also might be semantically deferring to LLM-based chatbots. The next natural question, if this is the case, is whether such deference is justified.

To summarize, LLMs are being integrated into our linguistic community in two dimensions: first, by providing a new group of experts responsible for defining the extension of a fundamental term, meaning; and second, by potentially transforming the network of semantic deference underpinning the linguistic community by introducing a new artificial speaker to whom non-expert speakers might be deferring. This picture of meaning, thus, explains how these language technologies are engaging in linguistic interaction, leading us to explore how successful these two dimensions of analysis are: the extension of meaning in LLMs and their role in the network of semantic difference. Fundamentally, both of these dimensions allow for the conceptualization of LLMs' modes of engagement in linguistic interactions without presupposing any capacity to simulate an inner process of meaning production. In this way, this framework offers a conceptual space to explore the problem of meaning in LLMs in a way that reconciles the intuition that these models are artificial speakers, despite being as "smart" as a toaster.

So far, I have summarized the theoretical contribution of this thesis. But why is it important to define a conceptual space that reconciles both intuitions from a non-theoretical point of view? As I remarked in Section 2.2.3, Floridi argues that cognitive approaches to AI agency risk misallocating efforts and resources toward an unrealistic and speculative ambition: replicating human intelligence. In contrast, by adopting a non-cognitive approach, we can avoid anthropomorphic fallacies and maintain reasonable expectations regarding the capabilities and limitations of these technologies. Thus, moving from conceptualizing The Problem grounded in cognitive to a non-cognitive notion is not merely a theoretical shift. As AI systems in general, and LLMs in particular, become increasingly embedded in various aspects of our daily lives, "recognizing their fundamental nature as unintelligent agents—rather than intelligent systems (which may also act as agents) is critical for their effective design, development, deployment, governance, and regulation, both ethically and legally" (p. 2 Floridi, 2025). Therefore, conceptualizing LLMs as artificial speakers, currently being integrated into our linguistic community despite their lack of a human-like cognitive capacity of meaning production, offers valuable insights into their potential and limitations, and helps develop a more suitable framework for their development and integration into human social structures and interactions (p. 3 Floridi, 2025).

Following Floridi, adopting a more straightforward explanation of AI's capacities and limitations, namely the MRA or ARA thesis, constitutes the most productive standpoint for designing technologies that contribute to the development of better societies. In the first chapter, I presented the point as the second advantage of adopting a non-cognitive framework for AI agency. In our specific case of study, adopting a conceptualization of LLMs as artificial speakers that engage in successful linguistic interaction in spite of being as "smart" as a toaster represents, in my view, the best stance from which to study the musts and must-nots of these technologies. In other words, in this thesis, I claim that a non-cognitive perspective on AI and, in particular, a non-cognitive and non-mentalistic approach to the problem of meaning in LLMs offers a more effective framework for examining the capacity of LLM-based technologies to be integrated into our linguistic communities

and guiding their impact in developing more just, diverse, and sustainable societies.

I conclude this thesis by suggesting potential future work based on the conceptual space this thesis opens. I consider that there are two fundamental research lines related to Chapters 1 and 2. First, I suggest a future research direction for those who are not convinced by my application of Putnam's Externalism in Chapter 2, but consider that it makes sense to contextualize the debate on the problem of meaning as a specific topic within the broader discussion of AI agency. In other words, I suggest a first future research line for those who are only convinced by the argument I developed in Chapter 1.

The future research line related to Chapter 1 involves exploring how other non-mentalist theories of meaning can be adopted to solve the problem of meaning, grounded in the acceptance of a non-cognitive notion of agency. I believe that different non-mentalist theories of meaning can provide valuable insights into the various ways in which these models are being integrated into our linguistic environments. Fundamentally, these new insights will arise if such theories are applied while being guided by a non-cognitive notion of agency. This means, broadly speaking, that one is not interested in whether LLMs encode, through their training, pragmatic norms, histories of use, or naming practices by successfully simulating the role of the speaker in the linguistic environment. Instead, from a non-cognitive perspective, one might ask how the integration of LLMs modifies a linguistic environment composed of pragmatic norms, histories of use, or naming practices. In this way, new and interesting ways of conceptualizing LLMs' modes of linguistic interaction might emerge from adopting non-mentalist theories of meaning, not to conceptualize the inner processes by which these models generate outcomes, but rather to understand how these technologies interact with the linguistic environment.

On the other hand, for those who are also convinced by the argument I endorsed in Chapter 2, a potential future research line consists of addressing the problem of meaning through a metasemantic and moderate application of Putnam's Externalism. Two possible routes emerge here: one for those more interested in how extensions are fixed. For those interested in this dimension of meaning, a possible future research line lies in investigating how the introduction of LLMs is modifying the extension of meaning by introducing a new area of expertise that defines it. Such a research line might involve, for example, reconstructing the technical term *meaning* as it is used in different paradigms within xAI techniques for LLMs.

Another research line emerges from examining how LLM technologies and, more importantly, LLM-based chatbots, are being integrated into the network of semantic deference in our linguistic communities. In terms of conceptualizing the potential transformative impact of these technologies, investigating and making evident how other speakers might be deferring to these new technologies regarding, for example, understanding technical terms, adopting a political viewpoint, or being informed about their own health diagnoses, might play a fundamental role. Finding ways to guide such processes of semantic deference and encouraging a critical stance toward these technologies is a crucial step in managing their potential transformative effects within our linguistic communities.

Bibliography

- Arkoudas, K. (2023). Chatgpt is no stochastic parrot. but it also claims that 1 is greater than 1. *Philosophy & Technology*, 36(3), 54 (cit. on p. 2).
- Austin, J. L. (1962). *How to do things with words*. Harvard university press. (Cit. on pp. 25, 33).
- Barry, D. (2025). Death by ai [Accessed: 2025-07-31]. <https://davebarry.substack.com/p/death-by-ai>. (Cit. on pp. 3, 5).
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623 (cit. on pp. 2, 4, 6, 10, 22, 39, 42, 43, 65).
- Bender, E. M., & Koller, A. (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:211029226> (cit. on pp. 4, 14, 21–26, 31, 39, 46, 47, 51).
- Bishop, J. M. (2021). Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11, 513474 (cit. on p. 3).
- Biswas, S. (2023). Chatgpt and the future of medical writing. (Cit. on pp. 2, 62).
- Blease, C., & Torous, J. (2023). Chatgpt and mental healthcare: Balancing benefits with risks of harms. *BMJ Mental Health*, 26(1). <https://doi.org/10.1136/bmjment-2023-300800> (cit. on p. 63).
- Bloch, A., & Shaul, R. (2024, May). Introducing the first intelligent ai knowledge agent in healthcare, using patients’ medical records for more accurate information and routing [Accessed: 2025-07-28]. <https://khealth.com/blog/about-k/introducing-first-ai-healthcare-knowledge-agent/>. (Cit. on pp. 2, 3, 62).
- Block, N. (1998). Semantics, conceptual role. *Routledge encyclopedia of philosophy*, 8, 652–657 (cit. on pp. 14, 15, 25, 29, 33).
- Boghossian, P. A. (2003). The normativity of content. *Philosophical Issues*, 13, 31–45 (cit. on p. 26).
- Borji, A. (2023). A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494* (cit. on p. 2).
- Bottazzi Grifoni, E., & Ferrario, R. (2025). The bewitching ai: The illusion of communication with large language models. *Philosophy & Technology*, 38(2), 61 (cit. on pp. 4, 14, 22–25, 33, 35).
- Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press. (Cit. on pp. 14, 24, 25, 35).

- Brandom, R. (2010). *Between saying and doing: Towards an analytic pragmatism*. Oxford University Press. (Cit. on pp. 14, 25, 35).
- Brandom, R. (2021). *Articulating reasons: An introduction to inferentialism*. Harvard University Press. (Cit. on pp. 14, 25, 35).
- Budding, C., & Zednik, C. (2024). Does explainable ai need cognitive models? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46 (cit. on pp. 59, 60).
- Burge, T. (1979). Individualism and the mental. *Midwest studies in philosophy*, 4(1), 73–121 (cit. on p. 61).
- Burge, T. (1986). Individualism and psychology. *The Philosophical Review*, 95(1), 3–45 (cit. on p. 61).
- Burge, T. (2007). *Foundations of mind* (Vol. 2). Oxford University Press. (Cit. on p. 61).
- Cappelen, H., & Dever, J. (2021). *Making ai intelligible: Philosophical foundations*. Oxford University Press. (Cit. on pp. 14, 21).
- Cascella, M., Montomoli, J., Bellini, V., & Bignami, E. (2023). Evaluating the feasibility of chatgpt in healthcare: An analysis of multiple clinical and research scenarios. *Journal of medical systems*, 47(1), 33 (cit. on pp. 2, 62).
- Christian, J. (2023). Amazing “jailbreak” bypasses chatgpt’s ethics safeguards [Accessed: 2025-07-28]. (Cit. on p. 2).
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (cit. on p. 2).
- Coelho, M., & Millière, R. (2023). The vector grounding problem. *arXiv preprint, arXiv:2304.01481* (cit. on p. 36).
- Dave, T., Athaluri, S. A., & Singh, S. (2023). Chatgpt in medicine: An overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6, 1169595. <https://doi.org/10.3389/frai.2023.1169595> (cit. on p. 62).
- Davidson, D. (1973). Radical interpretation. *Dialectica*, 313–328 (cit. on p. 34).
- Davidson, D., & LePore, E. (1986). A coherence theory of truth and knowledge. *Epistemology: an anthology*, 124–133 (cit. on p. 34).
- Dennett, D. C. (1971). Intentional systems. *The journal of philosophy*, 68(4), 87–106 (cit. on p. 34).
- Dennett, D. C. (1989). *The intentional stance*. MIT press. (Cit. on p. 34).
- Derosset, L. (2011). Reference and response. *Australasian Journal of Philosophy*, 89(1), 19–36 (cit. on p. 34).
- Devitt, M. (2021). *Overlooking conventions*. Springer. (Cit. on p. 34).
- Evans, G. (1982). *The varieties of reference*. Oxford University Press. (Cit. on pp. 25, 34).
- Floridi, L. (2023a). Ai as agency without intelligence: On chatgpt, large language models, and other generative models. *Philosophy & technology*, 36(1), 15 (cit. on pp. 3, 14, 21, 23).
- Floridi, L. (2023b). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press. (Cit. on pp. 3, 6, 13, 16–20).

- Floridi, L. (2025). Ai as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. *Philosophy & Technology*, 38(1), 30 (cit. on pp. 3, 6, 13, 16, 19–21, 67).
- Glüer, K., & Wikforss, Å. (2009). Against content normativity. *Mind*, 118(469), 31–70 (cit. on p. 26).
- Glüer, K., Wikforss, Å., & Ganapini, M. (2009). The normativity of meaning and content. *Stanford Encyclopedia of Philosophy* (cit. on p. 26).
- Grice, H. P. (1969). Utterer’s meaning and intention. *The philosophical review*, 78(2), 147–177 (cit. on p. 33).
- Grindrod, J. (2024). Large language models and linguistic intentionality. *Synthese*, 204(2), 71 (cit. on pp. 2, 4, 6, 14, 21–25, 27–29, 31, 33, 34, 45, 46).
- Gubelmann, R. (2024a). Large language models, agency, and why speech acts are beyond them (for now)—a kantian-cum-pragmatist case. *Philosophy & Technology*, 37(1), 32 (cit. on pp. 2, 5, 14, 22–25, 33, 35).
- Gubelmann, R. (2024b). Pragmatic norms are all you need—why the symbol grounding problem does not apply to llms. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11663–11678 (cit. on pp. 4, 6, 14, 15, 22–26, 31, 33).
- Haltaufderheide, J., & Ranisch, R. (2024). The ethics of chatgpt in medicine and healthcare: A systematic review on large language models (llms). *NPJ digital medicine*, 7(1), 183 (cit. on pp. 2, 62).
- Harnad, S. (1990). The symbol-grounding problem. *Physica D*, 42 (cit. on pp. 14, 21–23).
- Hlobil, U., & Bandom, R. B. (2024). *Reasons for logic, logic for reasons: Pragmatics, semantics, and conceptual roles*. Taylor & Francis. (Cit. on pp. 25, 35).
- Hua, Y., Liu, F., Yang, K., Li, Z., Na, H., Sheu, Y.-h., Zhou, P., Moran, L. V., Ananiadou, S., Clifton, D. A., et al. (2025). Large language models in mental health care: A scoping review. *Current Treatment Options in Psychiatry*, 12(1), 1–18 (cit. on p. 2).
- Kaplan, D. (1989). Afterthoughts. In J. P. J. Almog & H. Wettstein (Eds.), *Themes from kaplan* (pp. 565–614). Oxford University Press. (Cit. on p. 44).
- Kripke, S. A., et al. (1980). *Naming and necessity* (Vol. 217). Springer. (Cit. on p. 24).
- Kripke, S. A. (1980). *Naming and necessity: Lectures given to the princeton university philosophy colloquium* (D. Byrne & M. Kölbel, Eds.). Harvard University Press. (Cit. on pp. 34, 44).
- Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms. *Transactions of the Association for Computational Linguistics*, 12, 1087–1103 (cit. on pp. 4, 14, 22, 23, 33, 34).
- Lewis, D. (1970). General semantics. *Synthese*, 22(1–2), 18–67 (cit. on p. 44).
- Linzen, T., & Baroni, M. (2020). Syntactic structure from deep learning. *ArXiv, abs/2004.10827*. <https://api.semanticscholar.org/CorpusID:216080851> (cit. on p. 60).
- Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*. <https://doi.org/10.1016/j.tics.2024.01.011> (cit. on p. 60).

- Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *Computational Linguistics*, 50 (3). <https://direct.mit.edu/coli/article/50/3/1191/121670/Do-Language-Models-Words-Refer> (cit. on pp. 2, 4, 7, 10, 11, 14, 21–25, 33–35, 41–46, 49, 50, 55, 56, 60, 63).
- Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824* (cit. on p. 60).
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (cit. on p. 29).
- Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. MIT Press. (Cit. on pp. 25, 34).
- Millikan, R. G. (2004). *Varieties of meaning*. MIT Press. <http://ebookcentral.proquest.com/lib/reading/detail.action?docID=3338666>. (Cit. on pp. 25, 34).
- Millikan, R. G. (2005). *Language: A biological model*. Oxford University Press. (Cit. on pp. 25, 34).
- Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford University Press. (Cit. on p. 25).
- Mollo, D. C., & Millière, R. (2023). The vector grounding problem. *arXiv preprint arXiv:2304.01481* (cit. on pp. 4, 22, 24, 25, 30, 39).
- Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT press. (Cit. on p. 25).
- OpenAI. (2025, July). Introducing chatgpt agent: Bridging Research and Action [OpenAI blog post (product release)]. (Cit. on p. 3).
- Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381. <https://doi.org/10.1098/rsta.2022.0041> (cit. on p. 60).
- Pepp, J. (2025). Reference without intentions in large language models. *Inquiry*, 1–19 (cit. on pp. 14, 22–24, 31, 33, 34).
- Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of the association for computational linguistics: ACL 2023*, 13387–13434 (cit. on p. 2).
- Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *ArXiv, abs/2208.02957*. <https://api.semanticscholar.org/CorpusID:251371595> (cit. on pp. 22, 24–26, 29, 31, 33, 36, 48).
- Piantadosi, S. T., Muller, D. C., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, 28(9), 844–856 (cit. on pp. 4, 6, 14, 15, 22–24, 26, 30, 33, 36, 39).
- Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, 7, 131–193 (cit. on pp. 24, 25, 35, 43, 50, 52, 53, 60).
- Putnam, H. (1983). *Realism and reason: Philosophical papers, volume 3* (Vol. 3). Cambridge University Press. (Cit. on p. 54).

- Putnam, H. (1992). *Realism with a human face*. Harvard University Press. (Cit. on pp. 54, 55, 61).
- Putnam, H. (1994). *Words and life*. Harvard University Press. (Cit. on p. 54).
- Rau, A., Rau, S., Zoeller, D., Fink, A., Tran, H., Wilpert, C., & Russe, M. F. (2023). A context-based chatbot surpasses radiologists and generic chatgpt in following the acr appropriateness guidelines. *Radiology*, 308(1), e230970. <https://doi.org/10.1148/radiol.230970> (cit. on p. 62).
- Rudd, A. (1997). Two types of externalism. *The Philosophical Quarterly*, 47(189), 501–507 (cit. on pp. 53, 54).
- Rumbelow, J. (2023). Solidgoldmagikarp (plus, prompt generation) [Accessed: 2025-07-28]. (Cit. on p. 2).
- Searle, J. (1980). Minds and brains and programs. (Cit. on pp. 22, 33).
- Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge university press. (Cit. on pp. 24, 25).
- Shea, N. (2018). *Representation in cognitive science*. Oxford University Press. (Cit. on p. 25).
- Speaks, J. (2010). Theories of meaning. *The Stanford Encyclopedia of Philosophy (Winter 2024 Edition)*. <https://plato.stanford.edu/archives/win2024/entries/meaning/> (cit. on pp. 32–34, 41, 44, 48).
- Speaks, J. (2016). The role of speaker and hearer in the character of demonstratives. *Mind*, 125(498), 301–339 (cit. on p. 34).
- Stalnaker, R. (2017). Reference and necessity. *A Companion to the Philosophy of Language*, 902–919 (cit. on pp. 41, 44).
- Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M., & Zhang, W. (2024). History, development, and principles of large language models: An introductory survey. *AI and Ethics*, 1–17 (cit. on pp. 27, 28, 30, 60).
- Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. (2023). Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846* (cit. on p. 60).
- Weiss, G. (Ed.). (2013). *Multiagent systems* (2nd). The MIT Press. (Cit. on p. 20).
- Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Ed.). Wiley-Blackwell. (Cit. on pp. 25, 35).
- Wooldridge, M. J. (2009). *An introduction to multiagent systems* (2nd). John Wiley & Sons. (Cit. on p. 20).
- Wu, S., Shen, E. M., Badrinath, C., Ma, J., & Lakkaraju, H. (2023). Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339* (cit. on p. 60).
- Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4), 255–263 (cit. on pp. 2, 62, 63).
- Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2), 265–288 (cit. on p. 59).

- Zhang, M., Press, O., Merrill, W., Liu, A., & Smith, N. A. (2024). How language model hallucinations can snowball. *Proceedings of the 41st International Conference on Machine Learning* (cit. on p. 4).
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2025). Siren’s song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 1–45 (cit. on p. 2).
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38 (cit. on pp. 2, 27, 30, 59, 60).