Talking machines and linguistic cognition

A critical review of the use of large language models in linguistic theorizing

Alessia Giulimondi

August 27, 2024

Utrecht University

Research Master Linguistics

First reader:
Rick Nouwen

Second reader:
Tejaswini Deoskar

Acknowledgments

I would like to thank my supervisor, dr. Rick Nouwen, for the invaluable inputs he gave me. I was able to better understand my own ideas by being pushed to explain them further and in greater detail. I gained confidence in my capabilities with every meeting we had.

I would also like to thank Prof. Roberta d'Alessandro and Prof. Antal van den Bosch for their observations and feedback regarding the debate on innatism and LLMs. Their thoughts helped me frame and better understand the direction of my thesis and my ideas.

Finally, I am grateful for the support I received from dr. Olivia Guest and prof. Iris van Rooij, who inspired this entire research. Thank you for advising me and encouraging me while dealing with the wilderness of my thinking.

Abstract

Increasingly, linguistic studies are employing LLMs to explain the underlying mechanisms of human linguistic cognition by applying experimental methods to LLMS that were previously adopted to test human participants (Huebner et al., 2021; Beguš et al., 2023; Piantadosi, 2023; Goldstein et al., 2020). Computational cognitive scientists have argued that the assumptions underlying these research choices are incorrect(Guest and Martin, 2023; van Rooij et al., 2023) and a growing body of linguists is taking a critical stance towards LLMs (Martínez et al., 2023; Kodner et al., 2023; Katzir, 2023; Bender and Koller, 2020; Bender et al., 2021). However, meta-theoretical linguistic research is still scarce, and, so far, no systematic analysis of language studies using LLMs as experimental tools was conducted. This thesis aims to understand how the use of LLMs in research is affecting theory building in linguistics. More specifically, this analysis will focus on two research questions: 1) What is the theoretical relation of LLMs to human cognition, when they are used for linguistic research? 2) How valid is the use of LLMs in linguistic theory?

The thesis will review ten linguistic articles and argue that they share the assumption that LLMs represent an artificial replication of human linguistic cognition. Moreover, drawing from Guest and Martin (2023); Guest (2024) and Sullivan (2022) theoretical framework, it will be discussed how LLMs used to generate human-like linguistic behavior represent a theoretical misuse.

It will be shown how this misuse of LLMs is motivated by an industry-driven research mindset (Ahmed et al., 2023), which may be at the root of the theoretical misconceptions hypothesized in the first inquiry of this study.

This analysis is relevant for the understanding of language technology by language professionals and the possible systemic misinterpretations at play in research in a human-machine era.

Contents

1 Introduction

It is now common to begin research articles about large language models (LLMs) with acknowledgements of the remarkable progress deep learning research has made in the last few years, especially since the deployment to the public of Chat-GPT. Naveed et al. (2024) have shown that the growth of research related to LLMs both within and outside the field of machine learning is exponential, especially in the last two years.

When evaluating the advancements and shortcomings of deep learning from an academic perspective, one must observe that what made the great progress of deep learning accessible and appealing to the general public and to research fields outside of data and computer science was the important achievements in Natural Language Processing (NLP). As was observed by Harari (2023) what makes AI technology so special is its use of *language*, “the operating system of human civilization”. Harari suggests that these new technologies that go under the name of Artificial Intelligence (AI) are dangerous for human civilization because of their ability to reproduce human language in conversations with human users. Harari highlights how language represents the foundations of our civilization, as it would be unthinkable to build and preserve any society as we know it from historical accounts without the use of language. Thus, he argues that the danger of an artificial system that has learnt to manipulate linguistic items well enough to mislead even their own developers (see Matthews, 2022) is rather systemic, as AI is in principle able to access the communication system on which the institutions of our society runs. Moreover, he insists on the fact that AI represents an *alien* type of intelligence, the first one in history to be able to directly interact with humans efficiently and across large and diversified parts of the globe.

Mastery of language plays a crucial role in making these technologies accessible and useful for the general public, regardless of the fact that there is no real understanding of language achieved by these algorithms (Bender and Koller, 2020).

This particular focus on language, as the revolutionary aspect of AI systems, constitutes the starting point of this thesis, investigating the way linguistic research is approaching these *talking machines* and in which ways this is influencing linguistic theory. What is the role of LLMs in linguistic research? How do linguists use them and what are the underlying assumptions driving their studies?

In his famous article, Turing predicts that fifty years from when he was writing "the use of words and general educated opinion will have altered so much that one would be able to speak of machines thinking without expecting to be contradicted" (Turing, 1950, p.442). The quote foresees a world not so distant from his where the culture would allow the general public to discuss openly the cognitive abilities of machines, as they were humans. As recent debates about LLMs being sentient show (Matthews, 2022), Turing's prediction was impressively accurate. A famous illustration of this was the case of Blake Lamoine, an engineer fired by Google in 2022 after declaring that LaMDA (the LLM developed by Google) shows human-like consciousness (Wertheimer, 2022). Lamoine is not the only one who talked about consciousness of AI systems publicly (Auguera y Arcas, er 2). Interestingly, these claims come from specialists and not from lay people that one would imagine to be more inclined to such conditioning. Does this also have an effect on linguists, busy in the search for what makes humans speak and use language the way they do? In other words, how does the advent of artificial systems generating language change the scenario of research in language?

A great part of linguistics investigates language as a cognitive capacity with its rules and principles that may or may not be shared with other cognitive capacities (Hauser et al., 2002). In connection to the interest in the cognitive aspect of language, linguistics often shares theoretical frameworks and methodologies with the field of psychology. For this reason, the relation between language and consciousness is one of particular interest (see Schlinger, 2009, for a brief overview of the theory of mind and language acquisition). In this light, it becomes clearer why the turning point of our digital era is constituted by mastery of language of artificial systems, as language is what legitimizes talking about an Artificial Intelligence. That is, humans are able to guess and verify the consciousness of other humans only indirectly and only through their use of language. Language is undoubtedly central for our understanding of cognition and, even before that, for our intuition that a conscience exists. If this reading is not obvious, nor granted from Turing's words, Harari helps us to grasp this fundamental aspect of human language, by highlighting its centrality also for the history of human civilization. Without language the imitation game would not be so interesting for the public opinion nor for linguistics and psychology, but just for a niche of mathematicians and computer scientists. In other words, the imitation game, today clearly won by AI systems, is possible because computers can now fluently converse with us, not just because they can imitate the mistakes and hesitations of humans in some hard calculations.

Given this importance of language in AI, it is striking to witness the absence of linguistic theory in deep learning language modeling, as repeatedly pointed out by linguists (Kodner et al., 2023; Katzir, 2023; Backus et al., 2023; Baroni, 2022; Nature, 2023). The lack of linguistic theory and empirical linguistic evidence in the development of LLMs is well-renowned. For this reason a part of the current linguistic effort aims at building better LLMs or simply clarify what they can and cannot do (Bender and Koller, 2020; Kodner et al., 2023). However, if researchers from different fields, including developers themselves (Sculley et al., 2014), agree on the necessity of an improved AI, what "better" and "improve" actually mean is a complex matter.

The starting complexity of the matter is due to the fact that almost every scientific field is witnessing an AI race, that is, everyone has now the feeling that AI will finally answer long-standing questions that were untestable in the past. This situation is especially salient in linguistics, particularly the research programs that have a long history of rivalry (Valian, 2009; Tomasello, 2009) and are now turning to LLMs with the same questions, arguments and counterarguments. Section 3.1 will discuss this in detail.

Given the wide variety of applications of AI tools, of which LLMs represent only a small portion, it is useful to spend some words on the terminology generally used when referring to AI and how it will be used in this thesis.
As it is commonly used today, AI represents an umbrella term that contains in itself the old idea of recreating human intelligence in artificial systems, together with the attempt to understand human cognition (chapter 2 will elaborate more on this). Moreover, AI is at the same time a marketing strategy and a scientific field, as well as an organizational unit (see van Rooij et al., 2023, for a brief description of the history of AI and its multiple meanings). Thus, talk about AI and its impact on science and society is often confused. In fact, what is it really that we talk about when referring to AI? The AI systems central to this thesis are the ones that employ a deep artificial neural network to perform a specific task. Given that the task in question is language production, the focus of this analysis will be on LLMs, a subcategory of artificial neural network (hereafter, ANN) that is trained on linguistic input.

Within linguistic research using LLMs at least two types of studies can be identified: research that uses LLMs as a tool to conduct empirical investigations (Clark and Sakas, 2010; Huebner et al., 2021; Piantadosi, 2023; Oudeyer et al., 2019) and research that is interested in investigating LLMs' abilities and degree of (alleged) language understanding (Piantadosi, 2023; Baroni, 2022; Beguš et al., 2023; Seals and Shalin, 2023a,b).[1] In the former, LLMs are used to understand human linguistic competence, testing linguistic hypothesis on LLMs to explain better human linguistic behavior. The latter is more "engineering-oriented" and its main interest is testing LLMs' linguistic abilities to improve their performance. However, it is often the case that these two types of goals are pursued in parallel, supporting each other. The latter type is often involved in building models that look as similar as possible to humans. This approach often leads to affirming that human cognition has been replicated in computers, failing to distinguish between what it looks like human behavior and what is actually human behavior. Critiques to this type of research abound (Martínez et al., 2023; Katzir, 2023; Kodner et al., 2023; Guest and Martin, 2023) and they mostly come from the field of cognitive science, which highlight the intractability of recreating human cognition in artificial systems (Guest and Martin, 2023; van Rooij et al., 2023; van der Gun and Guest, 2024). To do so, they build a mathematical model to show the practical non-feasibility of research programs that aim at understanding (parts of) human cognition through the analysis of ANNs' behaviors and mechanisms.

It is uncontroversial that language is a fundamental part of human cognition, and this thesis intends to investigate whether approaches to ANNs in the computational cognitive neuroscience (CCN) literature reported by Guest and Martin (2021) and van Rooij et al. (2023) are also observable in emerging linguistic research. That is, CCN has a tendency to consider ANNs to be replicas of human cognition, as shown in Guest and Martin (2023). Given the strong relations between cognitive science and linguistics, it is important to analyze whether similar approaches are present also in a part of the linguistic literature.

To do this, I will investigate the theoretical assumptions that drive an emerging type of research in linguistics, which employs LLMs to explain human linguistic cognition. Specifically, the research question driving this study regards the way the use of LLMs in linguistic research is affecting theory building. The question aims at seeking the often neglected relationship between our scientific understanding of the world and the technological tools we adopt to gain this understanding. We often confidently rely on the conviction that technology is neutral (van der Gun and Guest, 2024) so that

[1] More types of applications and use are possible and certainly present in linguistics. For example, LLMs could be used for corpus analysis to gain insights about language performance (see as an example Grindrod, 2024, for the scientific utility of LLMs in sociolinguistics) Thus, the two types of research goals outlined here are not intended to be exhaustive, but it represents a useful categorization for the purpose of this analysis

our means do not necessarily define are goals. However, our devices are far from any neutrality, as it is known how they are developed in part outside of academic environments, responding mainly to necessities of the market (Ferrari et al., 2023; Ahmed et al., 2023). Their (in part) corporate nature is, indeed, undeniable nor hidden. It is, thus, perhaps non-trivial to investigate how these tools are influencing the way linguists conceive and research language as a cognitive and human phenomenon (Nature, 2023; Baroni, 2022; Goldstein et al., 2020). As part of the wider effort reported above to identify the ambiguities of AI affecting cognitive science (van Rooij et al., 2023; Guest and Martin, 2023; Rich et al., 2021; van der Gun and Guest, 2024), this research focuses on unpacking the relationship between the use of LLMs and the understanding of language as a part of human cognition in linguistic research.

Furthermore, it will be discussed how the influence of LLMs in linguistic theory is part of a larger influence that private corporations have on contemporary society (van Dijck et al., 2023; van Dijck, 2021; Couldry and Mejias, 2019; Kerssens and van Dijck, 2022; Ahmed et al., 2023). The deployment to the public of tools such as the one that goes under the label of AI are often hard to regulate van Dijck et al. (2023), and they too often reveal to be harmful for individuals, societies and the environment (Weidinger et al., 2022; Andrić and Kasirzadeh, 2023; Bender et al., 2021). Thus, a closer scrutiny of the potential harm that these tools may cause also to science is needed, and this review goes in this direction.

This thesis builds upon an in-depth analysis of ten research articles from different fields of linguistics to highlight emerging practices in language studies that can be related to the ones reported in the cognitive sciences. Chapter 2 will walk the reader through the theoretical framework adopted in this analysis, introducing meta-theory (2.1) and its theoretical tools, such as the basics of logical inference (2.2). The principle of multiple realizability will be explained (2.3), followed by the definition of a model (2.4) as it is adopted in this thesis.

After introducing the research questions and hypotheses (2.5), chapter 3 will describe the full analysis, reporting relevant citations to identify what theoretical understanding of LLMs in linguistic research is operating in the reviewed articles. Chapter 4 will address the validity of the use of LLMs in the reviewed publications, connecting this use to the digital ecosystem in which LLMs are embedded and the consequences for academic research.

2 Theoretical framework

2.1 The need for a meta-theoretical framework

There are deep connections between the field of computational cognitive neuroscience (CCN) and AI as a field (van Rooij et al., 2023). van Rooij et al. (2023) describe the birth of studies in artificial intelligence as embedded in cognitive science research. On the other hand, in Newell and Simon (1976) is also visible how already in the '70s the use of computers as tools for empirical investigations regarding human intelligence was central for the development of computer science. Thus, cognitive science and computer science shared the fundamental aspiration to understand intelligence, influencing each other with their framework, tools and methodologies.

van Rooij et al. (2023) describe how back in 70' (and perhaps even earlier), there was a fruitful interdependence between computer science and cognitive science. That is, a branch of psychology understood human cognition as a sequence of computations, which is now referred to as *computationalism* (Newell and Simon, 1976), and it allowed computational modeling run on computer programs to become a productive theoretical tool to investigate human cognition. Therefore, a part of AI can be considered in its origin a branch of cognitive psychology, involved in the development of cognitive models based on psychological knowledge.

However, this virtuous relationship weakened throughout the years. What became more interesting to invest in was performance-oriented and product-oriented research, propelled by the economic interest of private corporations (Ahmed et al., 2023), which are already a the forefront of the current digital revolution (van Dijck et al., 2023; van Dijck, 2021) in which AI technologies are embedded.

Nevertheless, an important role in this shift of interest regarding AI tools was played by computationalism, which made possible to conceive of machines as good hosts for recreating intelligence. van Rooij et al. (2023) refer to this with the term *AI-as-engineering*: the idea that human intelligence can be recreated in artificial systems, given that the human mind is an information-processing system that performs computations. Thus, computers were soon understood as human-like working devices, the study of which would have yielded useful results to explain human cognition. It is important here to highlight that computationalism does not necessarily imply the possibility of the recreation of intelligence in artificial systems. Understanding cognition *theoretically* as a computing systems, does not mean that computations hypothesized to be happening in the human mind can be reproduced in computers. The link between computers and the human mind is only theoretical and its scientific usefulness remains in the domain of theory. Thus, AI-as-engineering is an evolution of computationalism and not necessarily a fruitful one. van Rooij et al. (2023) show the intractability of this evolution, by building a theorem that proves how recreating cognition in computers is mathematically unfeasible. They imagine an engineer that under highly idealized conditions (e.g. unbiased data and access to any possible machine learning method, past and future ones) sets up to create Artificial General Intelligence (AGI). However, the project is doomed to failure as the replication of real-world, human level capacities (e.g. memory, vision, language) “will necessarily consume an astronomical amount of resources (such as time and number of samples)” (van Rooij et al., 2023, p.8). As this is true for AGI, they explain, it is also true for sub-parts of human cognition. The intractability of recreating human-like cognition is rooted in the high complexity of its real-world realization together with the still insufficient knowledge we have about it (see also Guest and Martin, 2023).

This brief history traced in van Rooij et al. (2023) provides an important perspective to any analysis that aims at uncovering the underlying motives and theoretical assumptions of cognitive sciences now using AI tools in their research. The strong connections between linguistics and cognitive science allows linguists to benefit from research in CCN to understand better the methodological practices present in linguistic research (for instance Martínez et al., 2023; Kodner et al., 2023).

van Rooij et al. (2023) and Guest and Martin (2023) provide a useful theoretical framework to investigate what is the relation assumed by linguists between LLMs and linguistic cognition. For this purpose, they define the principles of meta-theoretical thinking, to define useful guidelines for cognitive scientists. Guest (2024) explains how meta-theoretical awareness is key for a healthy research program and meta-theory works through the adoption of a transparent thinking, meant to ensure replicability in research. A transparent thinking is achieved through the use of metatheoretical calculi defined as "a (semi)- formal system to describe, define, and constrain the process of adjudicating over theories in a given (sub)field" (Guest, 2024, p.4).

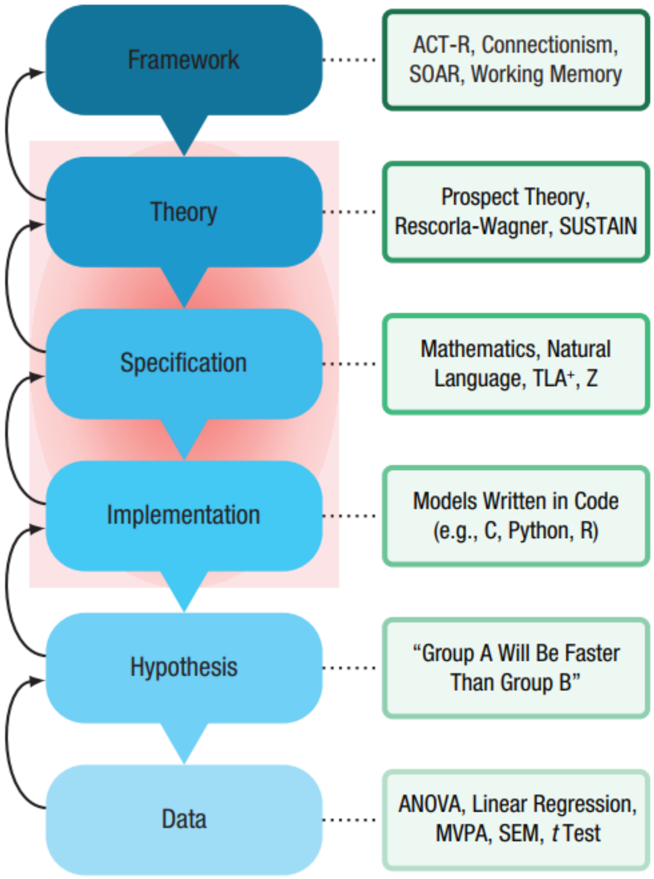
Framework
ACT-R, Connectionism, SOAR, Working Memory
Theory
Prospect Theory, Rescorla-Wagner, SUSTAIN
Specification
Mathematics, Natural Language, TLA+, Z
Implementation
Models Written in Code (e.g., C, Python, R)
Hypothesis
"Group A Will Be Faster Than Group B"
Data
ANOVA, Linear Regression, MVPA, SEM, t Test

Figure 2.1: One of the possible examples of a meta-theoretical model as proposed in Guest and Martin (2021)

In other words, a meta-theoretical calculus is a model of the way the scientist thinks, meant to increase the accessibility of the theory for both the authors and receivers of that theory. They are useful tools to define each step of our own reasoning, in order to identify flaws and jumps in the connections between the parts of our theoretical proposal and between the theoretical proposal, and the methodology and results. Specifically, what this framework proposes is a constrained path through which the connection between the theory and data has to pass in order to have robust explanations for the target phenomenon. Guest and Martin (2021) argue that too often cognitive science avoids the important steps of stating a theory with its specification and implementations (see Figure 2.1). Researchers should start by verbalizing the description of their theory. That is, they should explicitly define what is the phenomenon they are looking at, as in *what* the phenomenon *does* and then build a formal description of it, using either mathematical language or flowcharts. Then, an important step is an “executable implementation written in programming code” (Guest and Martin, 2021, p.791) of that formal description. In this framework, this implementation phase is also referred to simply as computational modeling, the only tool that enables researchers to understand their own claims and make them intelligible and usable for other researchers.

Nonetheless, they report the frequent lack of formal theorizing or specification (the metatheoretical calculus in the narrow sense) and the implementation in cognitive science, making claims about human cognition often ambiguous and the studies hard to replicate (Guest and Martin, 2023).

Precisely because of this lack of meta-theoretical awareness in CCN, Guest and Martin (2023) highlight the logical flaws present in many studies using ANNs to explain human cognition. Without this explicit iteration through theory, implementation, hypothesis and methodology, CCN is often inclined to make claims about human cognition that are not tractable or simply not verifiable through replications of the studies. Moreover, van Rooij et al. (2023) highlighted the importance, in CCN, of the idea that the human cognitive system can be replicated in artificial devices. The influence of this idea and the ambiguities allowed by the lack of meta-theoretical calculi formed a fertile ground for questionable links drawn between ANNs and human brains and behaviors. Lack of meta-theoretical calculi allows views like AI-as-engineering to become popular in CCN as they can ignore the fact that recreating human cognition in artificial systems is intractable. Without a rigorous modeling of theories with respect to data and implementations, it is possible to affirm the impossible. This means that an increasing part of CCN can explain human behavior on the basis of ANNs output, assuming an overlap of the model with the human brain.

In the next section the logical inference used in this type of work will be described, together with Guest and Martin (2023) criticism of that type of reasoning. Together with the principle of multiple realizability (2.3), this will form the fundamental argument against the use of ANNs as models of human cognition, which will support the current analysis in the identification of the same trends of CCN in linguistics.

2.2 Logical fallacies

In their criticism of the use of ANNs in CCN, Guest and Martin (2023) describe the meta-theoretical decisions of cognitive scientists explaining cognition with ANNs as flawed logical inferences. They employ formal logic to specify the steps taken in the studies in question and they highlight their logical fallacies. The kind of research in question addresses questions regarding specific cognitive capacities such as vision or memory or problem-solving. A cognitive capacity often investigated using ANNs is vision, generally studied with tasks involving visual object recognition. ANNs provide an interesting testbed as they fail to recognize objects that are clearly distinguishable for humans.

Because of the lack of meta-theoretical calculi common in the field (2.1), van Rooij et al. (2023) point out the ambiguous status of these types of studies with respect to their theory and their underlying assumptions, making the job of a meta-theorist more challenging, as it has to engage in discerning the logical fallacies where no formalization of the theory is provided. Therefore, Guest and Martin (2023) engage in a formalization of the steps left implicit in the studies they cite using the rules of logical inference. Logical inference serves as a meta-theoretical calculus to interpret the “sentences as found in the wild, in the CCN literature” (Guest and Martin, 2023, p.2015).

They observe that an inference rule largely employed in the literature is *modus ponens* (MP), which they formalize as:

$$P \rightarrow Q, P \vdash Q \tag{2.1}$$

which is read as: if P then Q; P is true; therefore Q is true. Guest and Martin argue that the field of CNN applies MP as follows:

1. if the model correlates with human behavioural and /or neuroimaging data, then the model does what humans do. $(P \rightarrow Q)$
2. the model correlates with human data. (P)
3. therefore, the model does what humans do. $(\vdash Q)$

The correlation in these cases is measured simply with input/output pairs. Thus, the ANN is prompted to see whether it responds in a similar way to how humans respond to the same questions/stimuli. In the case of comparison with brain activity, internal states of the models are compared to neuroimaging data to see whether there are similarities in the patterns of activation. The authors take an example from studies about the vision capacity using models trained on visual stimuli and build the following syllogism:

1. If the model correlates with human classification on photorealistic stimuli, then the model is impervious to adversarial images. $(P \to Q)$

2. the model correlates with human classification on photorealistic stimuli.(P)

3. therefore, the model is impervious to adversarial images. $(\vdash Q)$

In this case, being impervious to adversarial images [1] is an example for a model *doing what humans do*, which means processing images in the same way humans do.
To take an example from the linguistic literature, authors may use the following syllogism to talk about human syntactic competence and LLMs:

1. if the model correlates with human intuitions about certain syntactic structures, then the model has the same syntactic competence that humans have. $(P \to Q)$

2. the model correlates with human intuitions. (P)

3. therefore, the model has the same syntactic competence that humans have. $(\vdash Q)$

However, the conditional of this logical inference is false as correlation is taken as the only criterion for identity. In other words, when two systems correlate in their behavior they are considered to be the same thing. This means that if the model fits the data, it is also understood as a good explanation for the data, which means that it is considered to instantiate the mechanisms that causes the target phenomenon (Figure 2.2). The fact that the model fits the data in the same way the human behavior fits it is taken as evidence that the model is, indeed, the human system. However, as pointed out already in Guest and Martin (2023), no one actually believes it if put it as explicitly. Nevertheless, CCN, as a field, seem to operate on this premises according to Guest and Martin (2023).

To further explain it, the MP outlined above is based on the inappropriate causality that considers a good prediction to be enough to explain e.g. human vision. Thus, if the model is able to approximate human level performance in visual object recognition tasks, then it means that the model uses the same cognitive mechanisms that humans have. For example, Guest and Martin (2023) report this type of phraseology that “leaves the door open for

[1] Adversarial images are pictures that are subtly altered to make the model misclassify the image. Humans are generally not sensitive to these small changes, while ANNs tend to classify an image display flowers as a cat.

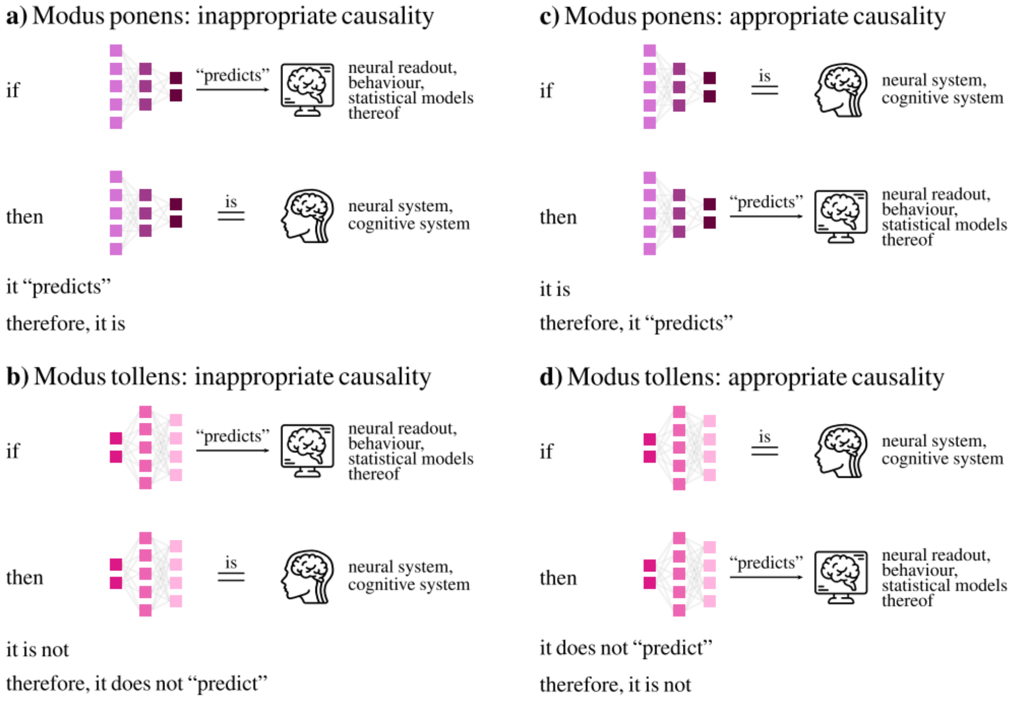
a) Modus ponens: inappropriate causality
if
"predicts"
neural readout, behaviour, statistical models thereof
then
is
neural system, cognitive system
it "predicts"
therefore, it is
b) Modus tollens: inappropriate causality
if
"predicts"
neural readout, behaviour, statistical models thereof
then
is
neural system, cognitive system
it is not
therefore, it does not "predict"
c) Modus ponens: appropriate causality
if
is
neural system, cognitive system
then
"predicts"
neural readout, behaviour, statistical models thereof
it is
therefore, it "predicts"
d) Modus tollens: appropriate causality
if
is
neural system, cognitive system
then
"predicts"
neural readout, behaviour, statistical models thereof
it does not "predict"
therefore, it is not

Figure 2.2: Visual representation of modus ponens (MP) and modus tollens (MT) as described in Guest and Martin (2023).

a logical fallacy": "[t]he fact that recognizable features of stimulus images could be reconstructed with a simple linear model [what we have generalized to statement P in this paper] indicates that the *latent* space represents properties that are also represented in brain activity [Q].". (p. 2018). This quote, reported in Guest and Martin (2023), is taken from Seeliger et al. (2018) and it is taken as an example of an over-extension of correlating results between humans and ANNs to give a mechanistic explanation of human vision. The flaw in this implied causal relation stands in the fact that correlation does not imply causation, but rather the other way around. That is to say that a correct causal relation should say that the model is the phenomenon, thus it can predict its behavior: "the models resemble the phenomenon because they indeed somehow capture (our beliefs about) the essence of the phenomenon, and not vice versa, i.e. models are not capturing the essence because they are correlated with the phenomenon" (Guest and Martin, 2023, p.219).

One may rightly argue that scientists do not operate in a world in which they know already what the underlying mechanisms that explain a certain phenomenon are (e.g. language, language acquisition), thus there is no possibility to use MP in the correct way. Cognitive scientists as well as linguists cannot yet affirm that the model reproduces the human cognitive (linguistic) system, because there is not yet agreement in the field about which cognitive mechanisms best explain language or memory. The field's primary goal is, indeed, finding those explanations. However, inverting the direction of the logical inference only seems to complicate the picture. As it is posed by Guest and Martin (2023), "we may simply not be in a state of affairs" where "we have a selection of good explanations to work from" (Guest and Martin, 2023, p.2023).

In their analysis, Guest and Martin (2023) also report another type of inconsistency identified in CCN literature regarding MP. One would think that when adopting MP, authors would as well accept *modus tollens* (MT) in case their results do not align with their predictions. Thus, if MP would lead to affirm that when the model and the human behavior correlate, then the model is running on the same cognitive mechanisms that humans have, it should also be the case that when the model does not correlate with human behavior, then the models differs from humans in its underlying mechanisms. This is called *modus tollens* (MT) and it is formalized as:

$$P \rightarrow Q, \neg Q \vdash \neg P \tag{2.2}$$

which is read as: if P then Q, Q is false; therefore, P is false.

To take again the example from visual object recognition, one can write the following syllogism following MT:

1. If the model correlates with human classification on photorealistic stimuli, then the model is impervious to adversarial images. ($P \rightarrow Q$)
2. The model is not impervious to adversarial images ($\neg Q$)
3. Therefore, the model does not correlate with human classification on photorealistic stimuli. ($\vdash \neg P$)

Notwithstanding, what CCN literature does is to simply affirm that the model needs to be further aligned with humans to finally show this correlation or that an updated version of the model will be a better representation of the human cognitive mechanisms (see Guest and Martin (2023) for a list of examples from the CCN literature). Therefore, after embracing MP in the outset of their research, researchers in CCN tend to disregard its logical inverse (MT) when finding results that contradict their hypothesis.

Interestingly, van Rooij et al. (2023) point out that it is often the case that ANNs fail to approximate human behavior, thus disproving researchers predictions about the models processing e.g. visual stimuli as humans do. According to them, this is unsurprising as it is mathematically unfeasible to recreate human cognition or parts of it (e.g. vision or language) in artificial systems, but the field still holds on to the idea that more data, or more "natural intelligence incorporated in these systems" (Guest and Martin, 2023, p.217) will eventually make a model able to reproduce human-like output, thus replicating its cognitive mechanisms.

2.3 The Principle of Multiple Realizability

An important aspect around which many criticisms of using ANNs to explain parts of cognition revolves is the principle of multiple realizability (Martínez et al., 2023; Kodner et al., 2023; Katzir, 2023; Warstadt and Bowman, 2022). The principle elucidates the relation between the *explanandum* and the *explanans*, respectively the object that has to be explained, hence the target phenomenon (e.g. language), and the construct that we use to explain it, hence the model, or simply our explanation. In the next section it will be explained what a model is and what it is not in the scientific tradition, and it will be clear from the start that the model does not have to be identified with the target phenomenon under investigation. This important dis-identification brings us to ask ourselves what is the relation between our models, our explanations and what we aim to understand. Meta-theory helps us to grasp this relation. Thus, one first realization is that a surface behavior of any kind can be due to a large variety of different underlying mechanisms.
Guest and Martin (2023) use the example of two clocks telling the time to describe the principle. One can learn the time of the day by looking at a clockwork clock and a digital clock and both will show the same time in the same way. However, the two clocks run on very different mechanisms, despite showing the same surface behavior. By applying the same logic outlined in 2.2, one can set out to discover the mechanisms causing the clockwork clock by looking at the behavior of the digital clock. Basing our conclusions on the correlations between the surface behavior of the digital clock with the clockwork clock we can be confident to say that the clockwork clock runs on digital mechanisms since this is what we know to be true for the digital clock.

This metaphor is very helpful to understand the logical fallacies often employed when studying cognition. Whenever researchers engage in investigations around the behavior of an object taken to be a plausible instantiation of (part of) human cognition, it is important to keep in mind the many possible explanations that can cause an identical behavior. Whatever we take our model to be, it is important to maintain a clear distinction between our model and the target phenomenon, in order not to ascribe unwarranted mechanisms to the phenomenon on the basis of correlating behavioral data.

The principle of multiple realizability enables our meta-theoretical analysis to recognize instances where researchers seem to consider models as pure replications of human cognition. When researchers disregard the principle of multiple realizability, combined with applying MP but not MT when data do not correlate with humans, they are treating the model as a replication of the target phenomenon, both in its behavior and in its underlying processes. Moreover, when the accessible mechanisms of the model are connected to neural activities in the human brain to explain surface behavior, the model is considered a replication of the human brain. These cases will be reviewed in this thesis to find evidence for this application of ANNs as models of human linguistic cognition. This application will be called "the replication assumption". A researcher is said to adopt the replication assumption whenever their research displays some or all of the following patterns:

1. False conditional premise in MP: prediction does not imply identity. That is, identical behavior in the models and humans does not assure that the model is human (i.e. the model's mechanisms are the same as the human's mechanisms).

2. Failure to apply the principle of multiple realizability. When the model successfully approximates human behavior, they assume that same mechanisms are causing the behavior in both humans and models.

3. Failure to apply MT after applying MP. That is, if the model does not show identical behavior researchers do not exclude all together the hypothesis that the model's mechanisms are the same as the human's mechanisms.

Point 1. and 2. describe the same failure taken from different perspectives. The first perspective identifies a fallacy in the conditional premise of the logical inference, analyzing it using logical form. The second one, shows the same type of fallacy applying a thesis belonging to the philosophy of mind tradition. They both reinforce each other in showing the inappropriateness of drawing conclusions from correlating data.

In this framework, the replication assumption is connected to the longstanding idea that human cognition can be recreated in artificial systems that share the same way of processing information, as hypothesized by *computationalism* (2.1). It is, thus, a natural evolution of this old idea to test models that are meant to generate human-like output to answer questions about the human cognitive system. Once a model is able to provide some testable behavior that equals humans in its performance, the belief of recreating intelligence in computers appears to be supported. Therefore, a strong tradition supports the replication hypothesis, even though not always explicitly stated as a hypothesis nor adopted as a framework. Indeed, it is rather an idea that interests the field of AI for historical reasons, as it is also shown by Turing's question on whether machines can think (Turing, 1950). Certainly, Turing did not intend to use the word thinking to refer to cognition. This reading is a rather modern one and does not belong to his view. However, the interest for a higher, complex problem-solving capacity and to what extent this was possible to be recreated in artificial systems was most probably an already fascinating thought. In other words, even though scientists do not explicitly say that they consider ANNs replications of human cognition, the driving force of the current renaissance of AI-as-engineering must be driven by Turing's same (unconfessed) fascination for this idea of replication.

2.4 What is a model anyway?

As already anticipated in the previous section, we will now turn to briefly define what a model is according to the introduced framework and according to general definitions of the scientific tradition. This will become useful in the analysis conducted in chapters 3 and 4.

van Rooij (2022) writes a concise but complete description of models in psychology. Their intention is to "bring conceptual clarity to what models ideally are in the context of scientific explanations", witnessing a lack of employment of scientific models in the field of psychology, despite the complexity of the mind and brain that it intends to investigate. Their definition of model in the context of cognitive science research will be adopted in the current study, integrating it with a standard definition (Treccani, 2003) and Sullivan (2022).

In van Rooij (2022), a scientific psychological model is defined as a "theoretical entity that psychologists use to understand and explain phenomena of interest". Moreover, a scientific model is also a type of representation, which is made of four parts: the object, the medium, the meaning and the user. The object could be, for instance, language, a part of human cognition linguists investigate, explained through a medium which can be a formal mathematical language as well as a flowchart. The meaning of the model is the content of the explanation which the model is helping to illustrate. Finally, the user is first and foremost the scientist him or herself as well as other scientists. Thus, the user is more generally the mind of the investigator that is aided in their process of understanding, together with the other investigators' minds, which can now rely on a transparent and unambiguous representation of what a potential explanation of the phenomenon might look like. As written by Guest and Martin (2021), "one of the core properties of models is that they allow us to "safely remove a theory from the brain of its author"(A. J. Wills, personal communication, May 19, 2020". (p. 2). Thus, models aid first and foremost communication among scientists, enabling them to test the plausibility and usefulness of theories: a tool for real open science (Guest and Martin, 2021). An important aspect highlighted by van Rooij (2022) is the "naive empiricism" that dominates the field of psychology which too often leads cognitive scientists to consider data descriptions as models with explanatory power. When referring to "naive empiricism", they define it as "the idea that scientific 'facts' can be derived from observations or empirical tests, independent of theoretical commitments".

A naive empiricist approach is one that easily leads researchers to attribute explanatory power to a model that can approximate human behavior, confusing explanations with good predictions, data with theory, as it was described in the previous sections. Thus, naive empiricism tend to leave all the work to the results of behavioral studies, without spending time in consolidating and testing their theories with an adequate formalization (see 2.1). The relation between the research hypothesis and the data, then, appears to be the only basis on which a theory holds, leaving aside the ocean of multiple explanations that may be actually hiding behind one statistically significant result. For instance, conducting multiple experiments to see how many words and what type of words are used when talking to children between 0 and 3 years old can help us to identify a specific type of communication happening with children and even compare it with the type of speech used with non-native speakers. However, it only gives us a behavioral picture of a particular communication setting and it does not necessarily give us any information about the underlying reasons for that particular behavior. Thus, if the studies are addressed with a behavioral research question, the relation with a general theory of linguistic competence or performance remains often unspecified. [2]

It is not surprising to witness this type of naivety when reading behavioral studies regarding linguistic performance, since linguistics is so embedded in cognitive science (see for an example McCoy et al. (2018) reviewed in 3.1).

[2] This is not intended to mean that such studies have no scientific value, nor that the research questions and the methodologies are invalid. It is merely a matter of perspectives: if cognitive science and linguistics aim at uncovering underlying mechanisms for certain surface behaviors, behavioral studies are crucial to support theories of cognition and language, as long as the paramount status of theory is recognized and it is actually used as a guidance in the wilderness of data and statistical analysis.

Thus, the lack of culture in the formalization of theories (e.g. by building a computational model to propose an explanation for language acquisition) seems to involve also the field of experimental linguistics, which usually does not formalize theories before running experiments. This leads language studies to the temptation of proposing theories only on the basis of narrations, often considering them as models. However, narrations are not models, even though they can certainly constitute a valid starting point to build one. According to the framework here adopted, computational modeling is the most efficient modeling practice to avoid ambiguities and under-specification of theories (Guest and Martin, 2021; van Rooij, 2022; van Rooij et al., 2023; Guest, 2024).

What remains unspecified in the definition proposed by van Rooij (2022) is the qualitative relation that exists between the model and the object under investigation. However, it is common to come across definitions that mention how models are simplifications of the object, like the one of the Oxford English Dictionary as reported in Börner et al. (2012). With a similar definition, Treccani (2003) proposes the following for *model*:[3]

> "a visualization of entities non representable in other ways, in their reduction to an empirical description, in the simulation of the logico-structural characteristics of a research object, via the creation of isomorphisms and analogies."

Even though the definition does not refer to the type of models using predicting algorithms, such as LLMs, what stands out from this definition is, again, the simplified, reduced status of the model in respect to the object. It is, indeed, clear that the model can only be an approximation of the object that it is meant to describe and thus it cannot be identified with the object itself. Thus, the model is a representation meant to help the conceptualization and communication of explanations in a formal and transparent way (van Rooij, 2022), which does not replicate in its complexity and completeness the real-world object. To borrow Guest and Martin (2021)'s words, (computational) models can be seen as "consumable scientific products on the way to deeper insights" (p.3). They are tools to aid understanding within our own minds and between minds, and they should not be confused with finalized products that replicate human phenomena, so that we can test them as they were humans. Thus, by keeping in mind the simplified status of our models, scientists can be more aware of their potential misuse.

[3] I translated the definition from Italian. I report the original here: "consiste nella visualizzazione di enti non altrimenti rappresentabili, nella loro riduzione a una descrizione empirica, nella simulazione delle caratteristiche logico-strutturali di un oggetto di ricerca mediante l'istituzione di isomorfismi o analogie."

Finally, the relation that stands between the model and the real-world phenomenon has been analyzed by Sullivan (2022), with important contributions specifically to the relation that ANNs have with real-world objects. In the next section, her perspective will be summarized as it will be useful to guide the analysis of 3.

2.4.1 How-actually models

Sullivan (2022) aims at demonstrating that the reason why ANNs cannot provide good explanations for real-world phenomena does not depend on the fact that they are black boxed. She explains how the researcher does not need to understand which calculations the model performed at each step to gain understanding of many phenomena. Of course, this might not be the case for cognitive scientists and linguists, which often address questions that require a detailed understanding of each level of implementation. Thus, she also clarifies that the level of black boxing can be an obstacle for explainability for target phenomena that require that level of detail in understanding the inner workings of a model.

Notwithstanding, Sullivan (2022) argues that in many cases it is the level of *link-uncertainty* that reduces or increases the explanatory power of a model. The link-uncertainty is defined as the missing connection between the model and the real-world phenomenon. This can be due to missing data, or data that may have different possible explanations and/or the field does not agree upon one explanation. The latter is, for instance, the case for cognitive science (Guest and Martin, 2023) and, consequently, linguistics. For instance, in the case of language acquisition, a long-standing debate revolves around the explanation for which children acquire their native language so fast and so accurately with no instructions and a restricted exposure to linguistic stimuli, both qualitatively and quantitatively. Many different explanations were proposed for this still puzzling phenomenon (Valian, 2009; Tomasello, 2009; Yang, 2004) and each one has enough supporting data to be scientifically convincing. This conflict makes the level of uncertainty in the field very high and, by consequence, the link between any model supporting one or the other hypothesis is equally possible, thus equally uncertain. Therefore, in the field of language acquisition the link uncertainty between the real world phenomenon of children acquiring their native language and the model proposed is inevitably high.

In order to explain the concept of link-uncertainty, Sullivan introduces the differentiation between *how-possibly models* and *how-actually models*. The former are models that describe a possible mechanism that explains a given real-world phenomenon, while the latter is the mechanism that *ac-*

tually causes (and, thus, explains) the real-world phenomenon. In order to understand better the difference between the two types, Sullivan takes a widely investigated example from the social sciences and philosophy of science: Schelling's model of segregation. This was a model originally built for the purpose of understanding the mechanisms that were causing segregation in American neighbourhoods. Schelling developed a simple algorithm, initially implemented on a chessboard, that was meant to simulate the dynamics that led to segregation in a neighbourhood. Two types of coins were used to act as the residents of the neighbourhood, one black and one white. Each actor-coin would have one simple preference: at least 30% of the neighbours is of the same type. Then, the algorithm works as follows: if more than 70% of the actors adjacent to a particular actor are of a different kind, move that actor to the closest unoccupied space.

The algorithm was then also implemented on computers and it was used to explain the segregating dynamics operating in real-world neighbourhoods. However, the problem with linking the model to the real-world segregation does not depend on how understandable are the steps taken in the algorithm, nor its implementation device - as we saw, the algorithm is as simple that it is possible to explain it in one sentence and it was originally implemented on a chessboard. What creates problems in drawing real-world, *actual* conclusions from this model is the theoretical assumption on which it operates. In other words, the problem is in the weak connections to real-world mechanisms that cause segregation that might not be identified in only one cause, namely one single preference of the residents. Intuitively, more factors contribute to the phenomenon of segregation in the real-world and neighbours preference may be one possible explanation among others. Thus, Schelling's model represents a good example of a how-possibly model, which was inappropriately used for years in the social sciences as a how-actually model. The inappropriateness lays on the high link-uncertainty between the model and the real-world causes of segregation.

This distinction is useful for this research since it clarifies the different roles that a model can play in the understanding of real-world phenomena. Moreover, it is particularly useful in research scenarios that employ ANNs, such as the ones reviewed in this thesis, as it represents a useful guidance in analysing the relationship between LLMs and the linguistic phenomena investigated.

It is clear from Sullivan's description of possible and actual explanations that in "real-world" research this distinction may not be so sharp and it may be present as a *continuum* where different levels of link uncertainty can be identified. It is also clear that how-possibly models have their valid scientific use. However, what is crucial for a healthy scientific study is not confusing how-possible explanations with actual explanations. In other words, what has to be clarified in studies that employ LLMs for linguistic research is their status, the type of explanation they are able to give, the added value of their adoption, given the explanations they can provide. Thus, where this distinction is not observable in the studies reviewed and how-possibly models are used as how-actually explanations, Sullivan's account provides a useful parameter to identify incongruities in the use of LLMs for linguistic research.

Moreover, it is important to clarify that the confusion between how-possibly and how-actually models is arguably preliminary to adopt the replication assumption. Thus, the identification of this confusion is helpful in understanding the logical inferences deployed in the studies. However, considering how-possibly explanations as how-actually explanations does not necessarily lead to the replication assumption, but, borrowing Guest and Martin's words, it leaves the door open for such a fallacy.

2.5 Research questions

This research aims to understand how the use of LLMs in research is affecting theory building in linguistics. This question entails an understanding of the status of LLMs in linguistic research and their relation to theory of language and research methodologies. What role do they play in experimental research? What are the metatheoretical choices that motivate the use of LMs in linguistic experiments? What is the metatheory that explains the use of LLMs in linguistic theory building?

Moreover, understanding the status of LLMs in linguistic research cannot be separated from the context in which these models are developed and deployed. Thus, understanding the relationship between who builds LLMs and the effects that these models produce in science is necessary. In the attempt to start answering these research inquiries, we can ask **how the use of language models is affecting linguistic theory** and break down this question into two smaller research questions.

RQ1 What is the theoretical relation of LLMs to human cognition, when they are used for linguistic research?

RQ2 How valid is the use of LLMs in linguistic theory?

Importantly, these research questions that will guide this analysis can only partially contribute to understand how LLMs are affecting linguistic theory. More questions can be formulated, as many factors operate in the choices linguists make when building their theories and adopting specific tools and methodologies. Socio-political and legal reasons as well as economic reasons can be behind the adoption of LLMs. Ethics and pure practicalities of research (e.g. LLMs are easier and faster to test than humans; they are readily available and do not need ethical approval) can motivate their use.

This thesis will take a specific linguistic angle, addressing the theoretical aspects that LLMs can influence. Thus, the use of LLMs is evaluated in respect to the kind of scientific knowledge that they foster. Legal, economic or ethical themes regarding the use of LLMs are here left to their respective fields.

Furthermore, it is also important to recall that changes and shifts that one can observe in linguistics, as well as in science more in general, are not attributable only to one cause or tool. The relationship between modern science and modern technology is old, complex and widely studied (van der Gun and Guest, 2024) and the way LLMs relate to linguistic research can have a place in the philosophy of science that investigates the mutual influence of science and technology.

3 Linguistics and the Language Model problem

When a technological tool is increasingly adopted in numerous research fields (Naveed et al., 2024) it is important that the methods and approaches surrounding the adoption of the tool are critically reviewed and assessed first and foremost within the field.

van Rooij et al. (2023) and Guest and Martin (2021) highlight the importance of internal evaluations of the way models and theories are used, without necessarily delegating meta-theorizing (see 2.1) to philosophy of science. Thus, cognitive scientists should develop their own tools to evaluate their meta-theoretical choices, since this internal evaluation is an integral part of the research.

In the same way, linguists should be able to develop meta-theoretical calculi to better understand their own research and their own choices at each step of the investigation process. With this internal, iterative, self-assessing course, the field would gain in quality and clarity of the research.

However, as it was observed for the field of cognitive science, the articles reviewed in this chapter lack explicit meta-theoretical explanations. This leaves the work of deducing what the connections are between the reference theory and the methodology to the reader, together with the assumptions that motivate the way the results are interpreted.

Therefore, the first research question precisely aims at assessing the level of meta-theoretical awareness in linguistic studies using LLMs, by uncovering any ambiguities present in these studies. This question intends to contribute to the understanding of the way LLMs are influencing linguistic research, by first reviewing what kind of reasoning and assumptions drive their employment in the linguistic literature.

As also stated in the previous chapters, the term LLMs refers to any artificial neural network (ANN) that is trained on text, regardless of the neural machine learning architecture (e.g. RNN, LSTM, transformer) and regardless of any additional non-neural training. Thus, it could also be possible to refer to these models as Neural Language Models, as they are language models with a neural network architecture. Finally, in all cases except one (Dong and Toneva, 2023) the models are trained on text, without any other type of input data (e.g. video, audio).

In chapter 2, it was explained how cognitive scientists misuse models, disregarding the principle of multiple realizability, and drawing inappropriate causality relations between the models and the human brain. Moreover, the problems of high link uncertainty between the models and the target phenomena were reported, showing how treating how-possibly models as how-actually models leads to unfounded conclusions about real-world phenomena and, thus, to unscientific explanations.
On the basis of this theoretical considerations, this analysis investigates the relations between LLMs and the studies' theoretical framework, hypothesis and results, to understand the logical inferences and meta-theoretical choices made in the research. It is, thus, analyzed whether similar logical patterns observed in the CCN' literature are also operating in a part of linguistic literature. There is a spectrum of possible different relations between LLMs and human linguistic behavior, ranging from LLMs being treated as how-actually models to being considered replicas of human linguistic cognition.
In this analysis, the term *replica* stands for an identification of the model with the investigated phenomenon (i.e. human language). When how-actually models are considered replicas of human linguistic cognition, the replication assumption is at play. As defined in chapter 2, the replication assumption involves the failure to apply the principle of multiple realizability and logical fallacies in the application of MP and MT.
LLMs are probabilistic models of language and, as any model as described in 2.4, they are an approximation of the target phenomenon. Thus, if one assumes an overlap of the two, this will result in a theoretical misuse and it can be recognized in studies that treat LLMs as replicas of an aspect of human cognition, such as language.

In order to investigate whether this misconception is at play, ten articles from diverse fields of experimental linguistics were selected, and statements and keywords from each article were collected to identify the authors' assumptions about the relation between the models and language cognition (see Appendix B and Appendix C). Precisely because the articles lack a meta-theoretical description, the analysis had to extract linguistic cues that enable the reader to understand the authors' assumptions, which are not always explicitly stated (see Appendix A for list of keywords).

The articles reviewed represent a random sample which is not intended to be a complete representation of the entire field of linguistics nor of the sub-fields they are part of. The only purpose of this analysis is to assess whether the same type of misconceptions that were observed in CCN (van Rooij et al., 2023; Guest and Martin, 2023; Rich et al., 2021) are represented in one of its sub-fields: experimental linguistics.

Therefore, the articles reviewed in this study are all engaged in investigations regarding linguistic cognition, whether they are interested in the acquisition of language or brain areas involved in language. This shows the deep connections between the field of cognitive science and linguistics, which can be also analyzed in light of their common assumptions to help research progress in the study of human cognition. Despite the fact that this sample may seem small, the restricted number of papers selected enables a deeper analysis of each study, from a meta-theoretical perspective.

The articles are mostly published with a reputable authorship, by distinguished journals. Their impact was judged based on the number of citations reported in Google Scholar and the majority of the publications have above 50 citations. The selection was done mainly starting from the reference list found in Guest and Martin (2023) and Huebner et al. (2021). Then, each paper had its own reference list from which I collected and selected more papers. This was specifically for the papers reviewed in 3.1. Regarding Dong and Toneva (2023), Galke and Raviv (2024) and Galke et al. (2024), I personally know the authors and I knew what they were working on.

Three main ways in which LLMs are used emerged in the analysis, thus the discussion of the results for this first research question will be divided into three main themes identified in the literature. The first, and probably most dominant, theme motivating the adoption of LLMs in linguistics is the chomskian hypothesis of language innatism, related to the poverty of the stimulus argument (Valian, 2009). The second theme regards brain areas involved in language and the way they work (Grodzinsky, 2000), referred to here as neurobiology of language. Finally, a third theme regards the way our knowledge of human language can help build better LLMs. Thus, this last theme is strictly connected to the role of linguists in this “AI summer” and the growing awareness of the importance of linguistics for the field of machine learning applied to language technology (Backus et al., 2023; Kodner et al., 2023; Seals and Shalin, 2023a,b; van Dijk et al., 2023).

3.1 The innatism dilemma

This section contains the review of four papers using LLMs to find evidence against language innatism. These are:

- Warstadt, A. and Bowman, S. R. (2024). *What Artificial Neural Networks Can Tell Us About Human Language Acquisition.*
- McCoy, R. T., Frank, R., and Linzen, T. (2018). *Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks.*
- Linzen, T. and Leonard, B. (2018). *Distinct patterns of syntactic agreement errors in recurrent networks and humans.*
- Chowdhury, S. A. and Zamparelli, R. (2018). *RNN Simulations of Grammaticality Judgments on Long-distance* Dependencies

The articles will be "organically" reviewed, reporting several citations as supporting evidence of the meta-theoretical choices of the authors. This means that the review reports the content of the papers together with the analysis provided in the current research about the interpretation of the authors' assumptions about the relation between LLMs and human linguistic cognition. The quotations will be put in relation to each other and not simply listed with a specific order. They will be grouped by themes and discussed as they were part of one single approach. This is intended to show the unified character of the underlying assumptions driving a certain part of the linguistic literature, highlighting the deep relations between the studies reviewed, both from the perspective of the type of terminology employed and the logical ambiguities allowed.

The most prominent theme observable in the literature adopting various kinds of LLMs is the one addressing the innatism debate historically central for any acquisitionists (Valian, 2009; Tomasello, 2009; White, 2020). The rapid advancement in machine learning of the past ten years has drawn attention of linguists involved in researching the fundamentals of language acquisition, leading them to revise traditional theories around which the majority of the debate was revolving (see e.g. Chomsky, 1988, for such a classical theory).

The theory of language innateness is usually attributed to the biolinguistic programme led by Noam Chomsky, where language is considered a biological organ, that should be investigated like any other organ of the human body, such as one studies a liver or the gall-bladder (Scheer, 2011). In this account, language is considered to be a computational system, a module of cognition, that can be boiled down to the capacity of recursion (Hauser et al., 2002). Thus, language, in its narrow sense, is a computational system and it evolved to create an "open-ended and limitless system of communication" (Hauser et al., 2002, p.1578). Moreover, another core aspect of Chomsky's research program is the fact that the language faculty is innate in every human being. This faculty, developed through 6 million years of human evolution, is now in the genetic makeup of our species and it distinguishes us from animals (Hauser et al., 2002). For this reason, children are born with a knowledge specifically tuned to acquire whichever language they will be exposed to between 0 and 6 years of age (see Fabbro, 2004, for an overview of the critical period in neurolinguistics). This innate knowledge is constituted by linguistic rules or principles that represent a guiding tool in the messy realm of the linguistic input children receive by their communicative environment. This hypothesis stems from the intuition that children must behave like a researcher that builds a theory of the system she is surrounded by, not merely blindly reproducing and memorizing occurrences and sequences of items. What children learn is to recognize patterns underlying the language, the structure that holds it together and they must be doing it guided by some innate flair for the specific structure that constitute the language faculty.

In this way, it is possible to see how the faculty of language is an organism that is then realized in each individual in a specific form given the language they are exposed to: the encounter of this innate linguistic capacity and the linguistic material creates what it is generally referred to as I-language (internal language) or, simply, linguistic *competence*. Competence is distinguished from the linguistic performance, which is the actual realization of language when it is used and it does not necessarily display the internal acquired knowledge, present in the competence. Therefore, a child may have in her competence at some stage of acquisition the knowledge of what a passive construction is, without ever producing it in her speech.

Nevertheless, if one accepts the metaphor of the child acting like a researcher, a problem related to the data arises as the input that children receive is qualitatively and quantitatively impoverished if compared to the more varied and complex ones adults receive (see Snow and Ferguson, 1977, for an overview of child-directed speech or Motherese). Moreover, the striking phenomenon of language acquisition is the ease and speed with which children acquire their first language without explicit instructions. Thus, according to the chomskian view, children must be endowed with some innate system, a pre-installed device, that fills the gap between the simplified version of the language they are exposed to and the complete and rich version they show to have in their performance at the age of 6 and in their competence even earlier in comprehension tasks (see Guasti, 2002, for an example of the acquisition of the Binding Principle). The simplified linguistic input children usually receive is generally referred to as the poverty of the stimulus, in the terminology adopted in chomskian frameworks. The poverty of the stimulus was for many years a key argument in support of the innateness of the faculty of language and it represented a central point around which most of the criticisms revolved around. Still in recent articles, who argues against innateness does so by falsifying the poverty of the stimulus argument, even though Chomsky does not seem to hold on to it anymore to discuss his theories. This is an important point to highlight here, as the literature dealing with chomskian research programs may often seem confusing. This is because many parts of the scientific debate, pillars of the research program, are often silently put aside, without explicit rejections or revisions of the previous theories. This often causes much research to remain focused on aspects of chomskian theories that are no longer so relevant. This appears to be the case for the poverty of the stimulus argument.

However, the interest of this thesis is not picking a side of the debate nor questioning the relevance of addressing innatism and its arguments. The only objective of this analysis is investigating the use of LLMs in linguistic research and the way this use is influencing the scientific understanding of human language. However, it is not a mere coincidence that a conspicuous number of the articles reviewed are specifically challenging the poverty of the stimulus argument and, consequently, the innateness of language. It seems reasonable to identify an aftermath of the "linguistic wars" (Harris, 2021) in this hot debate around LLMs and language acquisition (Kodner et al., 2023; Katzir, 2023; Piantadosi, 2023). Many opponents of the innatism hypothesis now see LLMs as tools that are able to provide a final proof for the implausibility of innate linguistic knowledge to explain first language acquisition. Thus, old experiments are conducted again with LLMs, already showing the unique status that LLMs have obtained in linguistic research. Even though consistent evidence was already found against innatism in decades of debates (as well as equally consistent evidence in favor of it), according to the "non-innatists", LLMs seem to add a stronger one, strong enough to re-open the debate around the poverty of the stimulus argument.

This brings us a to a final consideration regarding the nature of this debate. During the review process, it became clear that the authors of the articles about the poverty of the stimulus may be only concerned with using LLMs for an internal battle about the arguments employed by opposing research programs and not so much about the nature of language and its cognitive mechanisms. That is, it might occur to the reader (the one familiar with the debate about innatism and language acquisition) to think that LLMs do, indeed, a good job in supporting arguments against arguments and it seems sensible to use them to finally get rid of implausible, even though rather influential theories. I will elaborate more on this in section 3.5.

Despite the awareness that this may be the case, this analysis takes a step back from this type of discussions and considers all the articles reviewed as true scientific endeavours with the common objective of understanding language and its workings. Thus, only the scientific quality of the research in respect to our understanding of language is reviewed, leaving out the debate that might be the true underlying driving force of the writing.

3.2 LLMs and the innatism dilemma

Acquisitionists and syntacticians traditionally opposing innatism and the poverty of stimulus argument recognized in generative ANNs a powerful resource to prove the groundlessness of chomskian approaches (Piantadosi, 2023; Baroni, 2022; Huebner et al., 2021; Linzen and Baroni, 2021; Warstadt and Bowman, 2022; Chowdhury and Zamparelli, 2018; McCoy et al., 2018). A leading publication in this regard was Linzen et al. (2016), where a first theoretical and methodological framework was built to start reasoning against the poverty of the stimulus using Long-Short Memory Models (LSTM) probed for syntactic constructions. Baroni (2022) describes these new approaches as investigations “consisting in the experimental analysis of the grammatical skills of deep nets trained without the injection of any explicit linguistic knowledge” (p.2) and refers to it as “linguistically-oriented deep net analysis” (LODNA). Importantly, Baroni (2022) describes this type of analysis as “taking the perspective of a psycholinguist (Futrell et al., 2019), or perhaps more accurately that of an ethologist (McCloskey, 1991; Scholte, 2016), designing sophisticated experiments to “probe” the knowledge implicit in a species’ behavior” (p.2). Thus, LLMs are treated as a species behaviorally comparable to humans that it is worth investigating to understand human language. Interestingly, he complains how LODNA is not as common as it should be among linguists, despite its growing presence in annual events of the NLP and computational linguistics fields. This review, on the other hand, shows how LODNA and similar approaches are becoming more popular, showing important similarities in the way cognitive science research (Guest and Martin, 2023) and arguably many other research areas outside of linguistics (Aten, 2024) are using ANNs.

The deployment of Chat-GPT at the beginning of 2022 represented a crucial turning point in the language acquisition debate. The high performance of the AI-powered chatbot in conversations with humans provided strong additional evidence that it is possible to acquire language without assuming any innate linguistic knowledge (Piantadosi, 2023). In Piantadosi (2023) the main argument for this claim lies in the fact that an LLM like Chat-GPT can be defined as relatively unconstrained when exposed to the data and, without being instructed with explicit linguistic knowledge beforehand, they are able to generalize over the structure of sentences. Piantadosi argues this on the basis of the answers Chat-GPT gave when prompted with questions like "form ten sentences like 'colorless green ideas sleep furiously'". The article already received numerous responses and criticisms (Katzir, 2023; Kodner et al., 2023), thus it will not be part of this review, also considering that no experimental analysis was conducted nor any experimental work was systematically reviewed. However, it is important to report how explicit the literature can be in using these models to disprove theories that are highly influential in language acquisition, such as the poverty of the stimulus argument. Moreover, Piantadosi (2023) is exemplar in illustrating the ratio that is characteristic of all these type of studies: considering the model's output as evidence for cognitive mechanisms underlying the human language faculty. That is to say, the way Chat-GPT responds to human prompts has explanatory power of its underlying processes. In this case, human-level answers of the model imply human-like hierarchical generalizations over syntax and semantics. The surface behavior is, thus, caused by the same computational mechanisms, the same performance is due to the same type of competence. This study reports evidence that Piantadosi's approach is not unique, but rather systemic in a portion of the literature studying language as a cognitive capacity.

This analysis begins, then, with Warstadt and Bowman (2022) who report a large variety of experimental work using LLMs for questions regarding human language acquisition. The purpose of the research, as it is immediately clear by reading the title ("What Artificial Neural Networks Can Tell Us About Human Language Acquisition"), is explaining how LLMs can inform researchers about the way humans acquire language. The paper is extremely useful in outlining clearly what the purpose, hypothesis and methodological choices of this type of research are. They explain how the innatist theory can be researched in LLMs with ablation studies, which are not possible to conduct with human participants for ethical reasons. Thus, LLMs become extremely useful "artificial participants" free from ethical approval.

Ablation studies (or deprivation experiments) consist of removing from the input data direct evidence of the target knowledge. That is, LLMs are trained on textual linguistic input which lacks any occurrence of a particular hierarchical structure. Thus, the model's input is impoverished to make it comparable to the input that the poverty of stimulus argument assumes for children.

> "If the model succeeds after ablating A, it provides a proof of concept that the target is learnable without A. If the model furthermore does not enjoy any substantive advantages over humans,

> then we can conclude the result is likely to generalize to humans, and considerations from learnability do not justify the claim that humans require A.”

(Warstadt and Bowman, 2022, p.4)

When the authors refer to “substantive advantages over humans”, they are taking into account the major objection regarding this comparison between human learners and artificial learners (Katzir, 2023; Kodner et al., 2023), for which LLMs are exposed to an incomparably higher amount of data than children. This single acknowledgment challenges many attempts to draw direct connections between the models and the children as language learners. Therefore, the article reports how a lot of effort was spent in reducing the input to which the models were exposed to ensure the ecological validity of this comparison.
Huebner et al. (2021) focused precisely on this task, building a scaled-down masked LM based on RoBERTa, which they name BabyBERTa. RoBerta is a pre-trained model (Liu et al., 2019) that used the same architecture of BERT, Google’s LLMs deployed in October 2018 (Devlin et al., 2019). BabyBERTa is pre-trained with a more realistic input data to which an English-speaking 6 year old child is exposed. With this new model, the experiment aimed at building a “developmentally plausible dataset” on which LLMs can be trained to enable “claims about what children might learn without the aid of built-in linguistic knowledge” (p.632). One of the main objectives of the authors was to understand the role of child-directed language in language acquisition. Thus, LLMs are considered methodologically valid tools to explain language acquisition. The authors, indeed, claim that “although [LLMs] were developed for applications in language technology, their successes raise fundamental questions for acquisition research, including unsupervised grammar induction.” (p.624) and that their results “suggest that it is possible to acquire grammatical knowledge from substantially less data” (p.625). The authors refer to the results of the analysis they conducted to test BabyBerta’s grammatical knowledge. They employed the BLiMP grammar test (see Martínez et al., 2023, for a critical review of the BLiMP as used in this study), including only the words that were present in the training data, using a holistic scoring method because, as they reported “from a cognitive plausibility perspective, holistic scoring resembles much more closely the actual situation faced by humans tasked to judge grammatical acceptability.”(p.631).

These citations indirectly show authors’ assumptions about LLMs. Terms such as *cognitively* or *developmentally plausible* entail the assumption that these models are how-actually models of human linguistic cognition. However, the explanability of a model depends on its link-uncertainty (see chapter 2.4.1) and when this is high the model can only represent a how-possibly model, as in the case of LLMs in linguistic research. It is, indeed, often reported how these models learn language from a very different kind of source than humans. LLMs' linguistic inputs are in the great majority of cases only textual input decontextualized from any real, physical experience, which makes them incapable of learning actual meaning (Bender and Koller, 2020). That is, the lack of connection between the input and their referents in the real-world, which is granted to humans by their experience of the world, makes them impervious to meaning. Thus, LLMs are not learning language, as a system made, among other things, to structure and convey meaning. On the contrary, they are just learning the patterns of stimuli, not so differently from what they do with videos and images, of which meaning may not represent an intrinsic property for its complete recognition. For example, ANNs are outperforming dermatologists in identifying melanoma (Sullivan, 2022), a task that mainly relies on recognition of visual patterns from pictures. No meaning has to be retrieved from it. A binary response is sufficient to efficiently identify which image shows a melanoma and which one does not.

This aspect is only one of the multiple differences that drastically separate LLMs' mechanisms from human mechanisms of language acquisition (see below for more details, and Bender and Koller, 2020; Warstadt and Bowman, 2022). These considerations are often discussed to criticize the use of LLMs as "plausible" models for linguistic cognition (Katzir, 2023; Kodner et al., 2023; Bender and Koller, 2020). From this perspective, the mechanisms that a model uses becomes relevant for its utility for scientific understanding. This was a case that Sullivan (2022) considered but did not include in her definition of link-uncertainty. However, the mechanistic difference of LLMs and humans are also important to consider when evaluating the scientific appropriateness of their use.

In sum, if a high-link uncertainty can be observed in the relationship between LLMs and linguistic cognition, also an intrinsic difference between the human algorithms and the LLMs' algorithms are relevant to evaluate the explanatory power of these models.

These reflections on the incomparable nature of LLMs and human language acquisition are relevant especially when the literature shows a tendency to use them as participants comparable to humans for language acquisition experiments. Even though authors acknowledge LLMs' distance from human language acquisition in many aspects of the learning process, they still consider them usable replicas of human linguistic cognition. Indeed, in a related consideration, Warstadt and Bowman (2022) write "even imperfect models can provide useful evidence about human language learning.” (p.6).

What is crucial here to understand is that these studies draw conclusions about LLMs linguistic competence from their performance, exactly how it is done with humans. It is known, in fact, that LLMs often have a high level of black boxing, when, at the same time, performing human-like in linguistic tests (see as an example the way Karpathy, 2023, describes LLMs), and for this reason they may be more easily perceived as an odd kind of participant which mimics humans impressively well not only in its performance but also in its inaccessibility to its underlying mechanisms. Thus, it seems only natural to confuse them with replicas worth testing in the lab. However, as it was explained in chapter 2, the principle of multiple realizability should constrain researchers from drawing conclusions about human underlying mechanisms on the basis of correlations found in an artificial system’s behavior (i.e. LM). Nevertheless, the following quote clearly shows that Warstadt and Bowman (2022) do not hesitate to draw such conclusions:

> “although the input to humans is not annotated with linguistic features, training and testing models on a supervised task with such labeled data can still provide useful evidence about human learning (...) the experiment can tell us whether an inductive bias, such as a hierarchical bias or a compositionality bias, can be acquired through exposure to the unstructured learning environment.”

(p. 14)
and

> “model learners will be able to prove that specific linguistic behaviors are learnable under impoverished conditions, and thereby help to establish the causal roles of hypothesized advantages in the learning environment and the learner”.”

(p. 26).
From these claims, it is clear how a deep relationship between the way LLMs learn and the way humans learn is almost granted. What is most striking about Warstadt and Bowman (2022) ’s article is the full awareness, on the other hand, of the multiple limitations of their approach to which they dedicate the entire second half of their paper. First and foremost, they mention the data quantity problem already addressed in Huebner et al. (2021). That is, the fact that LLMs are trained on a disproportionate dataset, which does not resemble the real-world linguistic exposure of a child. Moreover, they also explain, as mentioned earlier, that the input is unimodal (in the majority of cases), as it is only constituted of text, without any sound or visual

input. Connected to this is the lack of prosodic information in the input and the absence of any kind of interactive learning, crucial for children in developmental stages (Bender and Koller, 2020; Logan et al., 2019). Thus, the authors explain how, respectively, LLMs fundamentally differ from human learners as they are mostly trained on written language data, while children learn by being exposed to sound (in cases of deaf children, they learn signs visually, when parents and/or caregivers are signers). They, then, explain how the input is mostly unimodal, lacking all the embodied cues that multimodal research has shown to be part of language and its acquisition (Perniss, 2018; Volterra et al., 2004; Butcher and Goldin-Meadow, 2000), being a crucial part of it prosodic cues and all factors that contribute to interactive learning (e.g. joint attention).

Despite this long list thoroughly explained, the authors keep affirming in the following section that "the less the model learner has an advantage over humans (independent of the experimental manipulation), the greater the chance a positive result from an ablation will generalize to humans." (p.20). Thus, the only pre-condition that has to be met in order to use LLMs to make generalizations about human linguistic competence is to reduce the models' size.

Given this important aspect of the reasoning, many studies - many of which are cited in Warstadt and Bowman (2022)'s overview - build their own LLMs with a LSTM learning technique and/or GRU RNNs (Chowdhury and Zamparelli, 2018; Linzen and Leonard, 2018; McCoy et al., 2018). This means that the only effort made toward LLMs that are more comparable to a situation faced by an infant acquiring her first language is reducing the input size and controlling exactly what is in the input. In this way, according to the authors, generalizations to humans are possible. However, all the others factors that divides LLMs from humans are disregarded, including the high-link uncertainty that makes them unsuitable from the start.

For instance, McCoy et al. (2018) used six different architectures of RNNs trained on 66 English words to test the models' hierarchical generalizations of subject-auxiliary inversion in English. This is a grammatical rule operating in English when a declarative sentence (1) is turned into a question (2):

1. The LM can speak.

2. Can the LM speak?

Both a hierarchical rule and a linear rule can explain this inversion, respectively "move the main verb's auxiliary to the front of the sentence" and "move the linearly first auxiliary to the front of the sentence". However, in complex sentences such as (3)

3. The LM that will help me can speak.

the two rules would yield different surface structures, as shown in (3a) and (3b):

3a. Can the LM that will help me speak?

3b. * Will the LM that help me can speak?

According to Chomsky (1971), as cited in McCoy et al. (2018), there are not enough examples in the input that children receive everyday to deduce the correct rule of subject-auxiliary inversion for complex structures. More specifically, child-directed speech is known to be simplified, with little use of complex structures such as the one in (3). Thus, only on the basis of sentences such as (1) and (2) children should conclude in favor of a linear structure. However, they are soon able to produce sentence like (3a) that presuppose a hierarchical generalization. Therefore, the innatist theory aims at explaining how children resolve around the hierarchical structure, despite most of the linguistic evidence to which they are exposed should lead them toward a linear structure. They argue that some innate hierarchical constraint must lead children to opt for a hierarchical rule.

The authors' main objective is to revisit this instance of the poverty of the stimulus argument by investigating whether a structure-sensitivity constraint is necessary to account for the generalizations that human language learners make.

In order to do so, they created a more realistic dataset on which they trained the model, to ensure a fair comparison with children's input. Then, they measured the subject-auxiliary inversion patterns of the models in question formation. In line with the reasoning outlined in Warstadt and Bowman (2022), they write:

> “ Of the six RNN architectures we explored, one of the architectures consistently learned a hierarchical generalization for question formation. This suggests that a learner's preference for hierarchy may arise from the hierarchical properties of the input, coupled with biases implicit in the network's computational architecture and learning procedure, without the need for pre-existing hierarchical constraints in the learner. We provide further evidence for the role of the hierarchical properties of the input by showing that adding syntactic agreement to the input increased the probability that a network would make hierarchical generalizations.”

(McCoy et al., 2018) (p.2). [1]
Thus, the authors draw conclusions about human internal syntactic representations by analyzing the output of an LM, as it is also clearly stated in the conclusion:

> "based on the behavior of one of the architectures we examined (GRU with attention), the answer to this question appears to be yes. The hierarchical behavior of this non-hierarchically-constrained architecture plausibly arose from the influence of hierarchical cues in the input."

(p.6).
The question of which the answer is yes is the one stated in the beginning of their article: "Is a structure-sensitivity constraint necessary to account for the generalizations that human language learners make?". Thus, they are affirming something regarding the internal mechanisms that are at play in humans on the basis of the LLMs behavior. Specifically, they claim that input alone is enough to explain the "hierarchical behavior" of the model, a result that is used to "revisit" the innateness hypothesis. *For the model* no innate knowledge for hierarchical structure is needed to acquire the constructions tested in the study and this must be the same *for humans*. Therefore, the principle of multiple realizability is ignored and LLMs behavior is used to explain human behavior. [2].

Furthermore, what is most interesting for this analysis is the authors' choice to base their conclusions only on positive results: "though the networks' common error types overlapped with the common error types for humans, the networks also frequently made some mistakes that humans never would" (p.5). In other words, the experimenters found mixed results, where the model did not always correlate with their human baseline, but they do not take this as evidence in favor of the innatist theory. This follows Warstadt and Bowman (2022)'s line of reasoning, for which "positive results from model learners are more meaningful than negative results." (p.4) and "positive results or proofs of concepts are more practically generalizable than negative results." (p.8). As explained in chapter 2 this same argumentation is also

[1] I assigned numbers to the article as pages are not numbered, not paragraphs

[2] They also "compare the networks' errors to the types of errors that humans make when acquiring English" (p. 5). The comparison is done between the only model that learnt the pattern in their experiment and results from Crain and Nakayama (1987) Thus, one first criticism regarding their methodology is the choice of comparing the models' output with human results to only one experiment that dates back to the 1987. Even though this may be seen as an independent issue, it is worth pointing out how the comparison results poorly designed.

present among cognitive scientists and it is criticized by Guest and Martin (2023) as misapplying the logical rules of inference using *modus ponens* and *modus tollens*. Considering only positive results when discussing the implications of the experiment fails to apply *modus tollens* ($P \to Q, \neg Q \vdash \neg P$), when the results do not support *modus ponens* ($P \to Q, P \vdash Q$). In other words, when the model fails to reproduce humans' behavior in all instances, authors choose to ignore them and conclude that the models behave human-like. This humanness in the behavior is, then, used to claim that the models are good explanations for human hierarchical representations (i.e. human cognitive representations of language). As explained in chapter 2, the problem with this type of reasoning is the way the conditional is posed and the conclusion is drawn from it. Similarly, Guest and Martin (2023) report the same flawed logic for explaining human vision:

> "When the models, in such adversarial cases, fail to classify images like a human, we do not conclude that this makes ANNs by definition unhuman-like(...) Instead, we tend to conclude that either the way the ANN has been trained, or otherwise designed, is dramatically different to humans."

(p.216).
The inferential failure is particularly clear in Chowdhury and Zamparelli (2018), where they explain in plain, clear prose the inferential process of *modus ponens* and *modus tollens* in the introduction:

> "if such a device [LM] could manage to ***replicate*** [my emphasis] fine-grained human intuitions inducing them from the raw training input this would be evidence that exposure to language structures (albeit in an amount orders of magnitude larger than the one a child receives, and without a connection to the non-linguistic context of utterance) should in principle be sufficient to derive a syntactic competence, against the innatist hypothesis (...) Suppose on the other hand that [ANNs] could approximate human intuitions on some linguistic phenomena but not on others, despite similar statistical distributions in the training input: this would now count as strong evidence that the 'unlearnable' phenomena tap on aspects of the grammar faculty that have limited representations in normal language samples, and are good candidates for being innate."

(p.134).
Their experiment is aimed at refuting the innatist theory, testing the models

on three types of syntactic structures that are used in the experimental literature on human subjects to test innatism. These are subject vs object relative clause; WH- extractions; subject and relative islands violations. The models are LSTM and a GRU RNN trained on a Wikipedia corpus. The models' output, prompted with a grammaticality judgement test, is compared to humans' grammaticality judgements reported in the psycholinguistic literature, under the assumption that the comparison "would have implications for the debate on language innatism" (p.134). The authors' questions investigate whether (R)NN are feasible models of innate-grammarfree language learners and which abstract properties they can learn from the input.
After running the analysis, the authors report that "this data shows that the increased perplexity with Wh cases has nothing to do with island effects (...)" (p.141) and that this "cast[s] doubts on the idea that the NN is using an abstract dimension of 'grammaticality'" (p.142).
In other words, results show many instances of non-human-like behavior and the authors conclude, in contrast with the logical inference described in their introduction, that

> "only after these issues have been resolved and a performance plateau has been reached we will be in a position to go back to the original question: are (R)NN feasible models of innate-grammarfree language learners? Which abstract properties can they learn from the input?"

(p.142).
In essence, the authors claim that if the model's behavior finally succeeds to align to the human behavior this would be proof of the non-existence of hypothesized mechanisms in humans' cognition (i.e. innatism). However, the logic is rather circular and essentially flawed, as RNNs might learn eventually to capture these structures with human intervention that might actually recreate the starting conditions that innatism theory claims to be present in humans - therefore at best supporting it.

To further understand the problems of the authors' reasoning, it is essential to recall the principle of multiple realizability and the impossibility of extracting any understanding of *actual* underlying mechanisms based on correlating behavior. Most importantly, unlike the example of the clock in 2.3, the results in both McCoy et al. (2018) and Chowdhury and Zamparelli (2018) do not show an identical behavior, but only a partial overlap. Thus, making any conclusions about what causes human language acquisition should not be granted. Furthermore, the circularity of Chowdhury and Zamparelli (2018)'s claims regards the ill-posed logical inference of MP and

MT, where MP should be the guiding inference of their study, but when results demonstrate a non-human-like behavior, authors fail to apply MT.

Moreover, they further add that

> "it could be tempting to take this [unhuman-like behavior] as a cue that even human ungrammaticality should be reduced to processing (...), but there are reasons to believe that, while processing might play a role, it cannot be the whole story."

(p.142).
In other words, the authors are clearly drawing conclusions about underlying mechanisms regarding human linguistic cognition, based on solely the models' behavior. This is supporting evidence that a clear overlap between the model and human language is assumed, to the point of considering LLMs as replicas of human linguistic cognition.
Similarly, McCoy et al. (2018) set out for a similar path, but conclude that

> " even if our findings do generalize to realistic language, we would only be able to conclude that it is possible to solve the task without a hierarchical constraint; humans certainly could have such an innate constraint despite it being unnecessary for this particular task."

(p.6).
Thus, at best, the model can serve as a how-possibly model for a specific language task, casting some shade over the relevance of these results for language acquisition. How-possible explanations against innateness already exists and they constitute a research program in itself (Tomasello, 2009). It is unclear what is the added value of testing LLMs on these tasks to arrive to the same conclusions. On the other hand, if one assumes that the models tested in these studies are considered replicas of human linguistic cognition, it becomes clearer how LLMs can contribute to a redefinition of the understanding of human language acquisition.

Linzen and Leonard (2018) find also a pattern similar to McCoy et al. (2018) and Chowdhury and Zamparelli (2018) in their results. They write that the human-like behavior displayed by the models "raises the possibility that the syntactic representations that emerge in RNNs are similar to those used by humans to process language." (p.2) [3]. Specifically, the authors tested human participants on a self-paced-reading test to see which agreement errors they made. They found three patterns of agreement error. First, errors are more likely in sentences like:

[3] I assigned myself the numbers, as the pages and paragraphs do not have numbers

> The sheer weight of all these figures make them harder to understand.

(based on Ronald Reagan, 13 October 1982; quoted in Wagers et al., 2009). These errors are due to the interfering presence of a plural noun that "attracts" the agreement of the verb *make*, which should agree with the singular subject *weight*. The second type of pattern shows that this agreement error is more likely to happen in sentences where the subject is singular and the attractor is plural, like in the example above. Finally, this error is more likely to happen in prepositional clauses (PCs) than relative clauses (RCs).

The simulation experiment consisted in testing an RNN on the same sentences to see whether the model showed the same agreement error patterns. Through the study of the RNN output, authors would deduce the syntactic representations of the models, which are assumed to be similar to the humans ones. That is, if the model shows a comparable behavior to the human one, the model (the RNN in this case) is considered to have the same syntactic representations that humans have for PCs and RCs. The error pattern in the agreement shows the hierarchical structure the model has. In other words, applying this logic equals to say that correlation implies identity and that one behavior can have only one explanation for its underlying mechanisms. This logic is observable from the following statement:

> "This [correlating behavior of the model with human behavior] raises the possibility that the syntactic representations that emerge in RNNs are similar to those used by humans to process language."

(Linzen and Leonard, 2018, p.2)
Again, underlying mechanisms of the human language faculty are hypothesized due to a shared surface behavior, disregarding the principle of multiple realizability. In the description of their results they write:

> "Two aspects of the networks' error patterns are consistent with the human data: first, agreement errors were more common when the local noun did not match the subject in number; and second, these attraction errors were more likely when the subject was singular and the local noun plural than the other way around. Unlike humans, errors were much more likely when the attractor was embedded inside a relative clause. (...) In other words, in RC-first sentences RNNs were similar to humans, but for the wrong reason: those sentences confounded proximity to the subject (the reason for human errors) with the presence of an attractor inside an RC modifier."

(p. 5 and 6).
One merit of this type of analysis is illustrating how the logical fallacy actually works in practice and finding that once the models are tested more accurately they show to drastically differ from humans. In this case, the authors first found a similarity to humans in the behavior of the model, but when they tested it again to specifically spot the way the model was handling the structure, they found that no hierarchical representation like the one used by humans was likely to be at play in the model. Thus, they showed how a similar behavior can be caused by different mechanisms, in this case, models show to rely on the linear structure of the sentence, while humans exploit hierarchical representations to parse sentences. However, in most cases human behavior and the models' behavior appears to be the same.

Given these conflicting results, the authors simply summarize their findings, without any discussion that meaningfully brings them together. They limit themselves to write: "these findings suggest that the syntactic representations acquired by RNNs differ from those used by humans in sentence processing.". Therefore, once again, the contribution of such a study remains understated, unless some special status is attributed to these models, the investigation of which would bring evidence for human language that no other model or technique was able to bring before.

To further explain this analysis, it is useful to go back to Warstadt and Bowman (2022), where the logic of these approaches are thoroughly reviewed and it is clear how its foundations are to be found in learnability studies (Gold, 1967). Briefly, research in learnability is concerned with investigating what is in principle learnable, using mathematics. Thus, computational models are built to assess what kind of data, distribution, learning constraints and learning environment make the learning possible. In line with this kind of research, LLMs seem to be treated like a new type of computational model that can inform linguists about the nature of learnable input. However, it is known how these models are built with very different purposes (Liu et al., 2024; CBS Mornings, 2023). Indeed, neural networks architectures were not built to propose new linguistic theories and they are far from resembling mathematical computations meant to understand the principles of learning. That is, their inner computations reveal little about learning mechanisms and the only reason why they are considered in research is their rather unexplained capacity to mimic human-level language. Nevertheless, even supposing that LLMs may reveal effective in showing the limits and conditions of language learning, they seem to be adding nothing new to what it is still possible to learn with traditional tools from learnability (Kodner et al., 2023). Therefore, at best, LLMs may be considered a redundant tool, often more laborious, expensive and resource-hungry than the traditional, non AI-powered means (Bender et al., 2021; Weidinger et al., 2022).

Finally, it is important to mention how the the principle of multiple realizibility is not unfamiliar to the proponents of the use of LLMs as good explanations for human linguistic cognition. Indeed, Warstadt and Bowman (2022) touch upon this problem, despite not proposing any viable resolution:

> " We can also study performance to make inferences about competence. We can construe performance very broadly to include many aspects of behavior, ranging from acceptability judgments to order of acquisition and reading time. Although this has its limitations—two systems that have identical behavior in some respects could have very different internal functioning—the more behavioral similarities we observe between two systems, the greater the evidence that they share an underlying mechanism."

(p.9)
Thus, despite describing the principle of multiple realizability, the authors still conclude that competence, understood as an underlying mechanism of the human cognition, can be uncovered studying the behavior of an artificial system, provided that the two systems share some behavioral similarities (i.e. performance). Moreover, they write in a footnote:

> "Dupre (2021) discusses the relation between ANNs and competence at length and suggests that ANNs are better viewed as models of human performance rather than competence because they are optimized to reproduce the output of human performance. We broadly agree with this view, and note that it does not contradict our claim that competence for ANNs may still be well-defined and testable."

(p.9).
However, it remains unexplained which principles allow us to make any realistic conclusion about human cognition based on the performance of a non-human system.

Considering the clear awareness of the problem (which they categorize as a pure limitation), the only reasoning that explains such an ambiguity is to deduce that the authors consider LLMs more than just simple models, but a replica of human linguistic cognition. Under this assumption, the choice of comparing the behaviors of two systems (one human and one artificial) would become more understandable. If we consider LLMs a testable replication of a human participant, analyzing their performance to find their competence is nothing different from what has been done so far in psycholinguistic research. Given that different individuals share approximately the

same cognitive skills, it is possible to compare their performance to derive one possible linguistic competence. [4]
Comparably, testing LLMs' performance against human performance to make predictions about human competence (together with LLMs' competence) may only come from the assumption that LLMs must be a replication of human linguistic cognition. Moreover, this assumption would explain the contradictions and fallacies highlighted in this section regarding MP and MT, the principle of multiple realizability and the distinction between how-actually and how-possibly models.

3.2.1 LLMs and language evolution

A field related to language acquisition is language evolution research, which is also interested in understanding the principles of learnability to investigate how language evolved throughout time (Kirby, 2002).

Recently, some research in language evolution has emerged that builds upon many studies reviewed here regarding the poverty of the stimulus debate. I will now review two articles (Galke and Raviv, 2024; Galke et al., 2024) to show the commonalities of language evolution research using LLMs with the studies reviewed in 3.2.
Language evolution is directly linked to language acquisition, and for this reason employing LLMs in a similar way to language acquisition studies helps researchers understand patterns of language emergence in humans by looking at what improves the models and what does not. Specifically, Galke et al. (2024) are interested in testing the role played by compositionality in language emergence and compare human responses to the output of GPT 3.2 and an RNN trained from scratch. Their research question investigates whether deep neural network models exhibit the same learning and generalization advantage when trained on more structured linguistic input as human adults. More precisely, their hypothesis predicts that a more structured language (i.e. a language system that displays more compositionality) will represent an advantage in language learning and, thus, it represents a milestone in the steps of language evolution. To investigate this hypothesis, they study whether both artificial learning systems (a LLM and a RNN trained from scratch) and humans benefit from a more structured language. They tested all three systems (humans, LLM and RNN) on ten different languages with different degrees of compositionality and their

[4] I leave aside for the purpose of this analysis the debate around performance and competence and whether these two are distinct aspect of language and cognition. For sake of simplicity, I here adopt the traditional distinction between competence and performance as also referred to by Warstadt and Bowman (2022)

> "results suggest that more structured languages are easier to learn, regardless of the learning system: a human, a recurrent neural network, or a large language model. Thus, generalization capabilities are heavily influenced by compositional structure, with both biological and artificial learning systems benefiting from more structured input by facilitating more systematic and transparent generalizations."

(p.12).
By reading this statement it seems clear that the authors are mostly interested in studying language taken as a separate, autonomous, natural object that can be observed independently of its embedding in human cognition (see also Grindrod, 2024). However, this is not explicitly stated in the article and it has to be regarded as a possible interpretation of the reader. Nevertheless, it seems important to point this out, as it would explain why the authors chose to test an LLM and an RNN together with humans.

However, what the authors indeed write is that their

> "findings strengthen the idea that language models are useful for studying human cognitive mechanisms, complementing the increasing evidence of similarity in language learning between humans and machines"

(p.11).
Thus, the bald connection identified in 3.2 is still operating here, suggesting the possibility of understanding human cognition also from a language evolution perspective by comparing human and artificial outputs. Moreover, they add that these "results predict that children would also benefit from more systematic compositional structure in the same way adults do – a prediction we are currently testing." (p. 11). Also in this case, models' behavior become predictors of human behavior, showing that authors conceive these models as how-actually models. Moreover, it is again unclear what type of an advantage LLMs may serve as for these questions, as adults and children can be directly compared (arguably more appropriately) to obtain some solid answer.

Additionally to these remarks, it is worth noting some degree of ambiguity in the intents as the authors state that their

> "work contributes to the understanding of deep neural networks and large language models, sheds new light on the similarity between humans and machines, and, consequently, opens up future directions of simulating the very emergence of language and linguistic structure with deep neural network agents."

(p.4).
It is, thus, clear how the objectives are bidirectional : one aims at understanding LLMs as an artificial construct using language and investigating what kind of linguistic input it can benefit from, while the other hints at a direct connection between the humans and machines that can help research gain understanding about language evolution.

In a related article of the same authors, (Galke and Raviv, 2024) they are more explicit in reiterating a tendency already described in 3.2, where improving the models' performance will increasingly justify its treatment as a how-actually model and, in most cases, as a replica of human cognition. This approach is, indeed, visible in the closure of Galke and Raviv (2024), where they write:

> " Integrating these biases into large language models may very well lead to more cognitively plausible models for gaining new insights on how children acquire their first language."

(p.12)
By reading this last statement, it is clear how the field of language evolution shows an emerging trend inclined to support the replication assumption, drawing from language acquisition the studies that work with it.

3.3 Neurobiology of language with LLMs

Another topic frequently studied in language using LLMs is the neurological basis of language, that is the study of the brain areas involved in language production and comprehension (Grodzinsky, 2000). This fields contributes to the understanding of language and human cognition in general by looking at the physiological aspect of it.

This field has identified a strong potential in the employment of machine learning in their methodologies, especially because the conceptualization of machine learning models draws from the neural structure of the brain, as the neural network architectures clearly show (Liu et al., 2024; Marcus, 2018; Sullivan, 2022; CBS Mornings, 2023). This has led neurolinguists to explore the possibility of understanding the brain by looking at models that generate human-like outputs and seem to be built to mimic some structures of the human brain (Dong and Toneva, 2023; Goldstein et al., 2020; Lakretz et al., 2020).

This section will, thus, review three studies (Dong and Toneva, 2023; Goldstein et al., 2020; Lakretz et al., 2020) that attempt to explain brain activity related to language through the analysis of LLMs' output and internal states.

Dong and Toneva (2023) conduct a meticulous investigation of the internal states (the activity of each cell in the neural network) of a multimodal video transformer, MERLOT reserve (Zellers et al., 2022), pre-trained on 20 million YouTube video, combined with text, and audio input. Thus, a multimodal neural network represents an attempt in the direction of a multimodal language framework that is interested in investigating how different modalities interact in language production and comprehension (Perniss, 2018; Holle and Gunter, 2007; McNeill, 1992). As also highlighted in Warstadt and Bowman (2022), a major challenge for linguists studying LLMs as models of human language is to make them comparable to the way humans use, perceive and interact with language. One dominant concern is, thus, the multimodality of language, highly connected to embodied cognition (Hostetter and Alibali, 2008), which represents a challenge to recreate in artificial agents. Thus, multimodal neural networks are conceived as a first step forward in the direction of more "cognitively plausible" models of human language (Warstadt and Bowman, 2022).

However, it is important to point out that the ways the term multimodality is used in machine learning and linguistics do not precisely coincide, as multimodal language studies are mainly driven by gesture research (McNeill, 1992; Holler and Beattie, 2003), adding on to it other body cues such as eye-gaze, facial expressions or speech errors (Slonimska et al., 2015; Clark, 2002; Holler, 2022).

On the other hand, multimodality in machine learning is mostly used to refer to models that are trained not only on text but also on video recordings images and audio (see for an example Tagliaferri et al., 2023), like in the case of Dong and Toneva (2023). Thus, Dong and Toneva (2023) are mainly interested in understanding the integration at the neural level of language and vision. [5]

Their research questions are stated as follows:

1. Do models learn any brain-relevant shared information between in-

[5] From this approach it is clear that the authors do not adopt a multimodal language framework, but treat language as a separate element from vision, as it is conceived and understood in machine learning. Here vision is simply the combination of images and audio to which the model is exposed to, in order to learn information that are otherwise harder to learn in a text-based input. On the other hand, when vision is involved in multimodal language studies, authors are mostly referring to it as iconicity and they are interested in the iconic properties of language (Slonimska et al., 2020, 2022; Perniss et al., 2010). Thus, vision is investigated as a property part of language itself and embedded naturally in our cognition.

dividual modalities, such that the joint representations [6] are better predictors of brain activity than the ones without visual stimuli?

2. Do models learn any brain-relevant new information when individual modalities interact, such that the joint representations are better predictors of brain activity than the sum of the ones without visual stimuli?

By looking at the first research question, it is immediately clear how the model may be easily understood as a how-actually model of the brain. If the underlying assumption is that good predictions make good explanations, these models are better models of the human brain activity just because they correlate in their internal states with the human brain activation of some language areas. The second research question, instead, is more oriented to use the human brain to improve the model, which is in line with the more traditional approaches of computer science research Newell and Simon (1976). They, write: "we turn to the only system that we have that truly integrates complex visual and complex language information– the human brain–to improve our understanding of vision-language interactions and integration in a popular multi-modal video transformer." (Dong and Toneva, 2023, p.1)

Therefore, their objectives are bidirectional: one goes in the direction of improving LLMs performance by using a "human derived-reference" that is able to "*capture*" the multimodal essence of the linguistic input to which humans are naturally exposed and one contributes to neurolinguistic research in demonstrating how vision plays a role in the human brain when processing language.

Thus, the authors analyze cell by cell the internal activity of the neural model to generalize over humans' brain activity when watching video recordings that activate both auditory and visual areas of the brain to process language. By comparing the fMRI results of 5 subjects watching fragments of the TV show *Friends* and the cell activity in each layer of the neural network trained on the same type of input, they identified a partial alignment with the human brain activity when exposed to multimodal input. Their partial findings are due to the fact that the most considerable alignment is to found in the later layers of the model, which "encode the most *brain-related* properties of video stimuli" (p.6). They further add that the integration of visual and linguistic inputs observed in these layers of the model are "*possibly akin* to the convergence of vision and language representations observed in the angular gyrus region." (p.6).

[6] representations are here intended to be the input presented to the model.

Given these statements, it is important to spend some words on the terminology used in the article, as they adopt the terms *brain-relevant* and *brain-related* throughout the paper to highlight the strict relation between the models' neural behavior and activity in the language areas in the human brain. As the reader may understand them, these terms are meant to cautiously advance a proposal for an understanding of LLMs as a plausible duplication of the human brain. However, the authors may be simply trying to find a term that highlights the similarities between the biological activity and the artificial activity of the neurons, without necessarily implying any kind of replication. Nevertheless, the intuitions and interpretations of results largely suggest the first reading. In support of this reading, is worth reporting two more citations where the authors write:

> " We suspect that the benefits of vision we identified by the ablation of vision information in the current models are far from fully encompassing the entire spectrum of multimodal integration processes taking place in the brain."

(p.6).
Thus, the models are conceived as (partially) replicating the brain activity from which conclusions about human language can be drawn. By this phrasing, the reader understands that the models are used to study the role of vision integration in language processing in the human brain, as they are acknowledging the possibility that other aspects of multimodality are at play in the brain other than visual object processing. In other words, these statements necessarily implies that internal states of an LM are good representations of brain activity when exposed to the same kind of stimuli of a human participant. Thus, the models are not only intended as how-actually models, as in good explanations of language processing, but also as a replication of the brain activity in an artificial system.

To show that this is not a one-case phenomenon, I included a literature review (Lakretz et al., 2020) on similar approaches in this study. Lakretz et al. (2020) engage in a thorough review of psycholinguistic literature regarding sentence processing and propose ANNs as more precise models to explain human cognitive mechanisms involved in sentence processing. Thus, the models are considered plausible implementations of cognitive mechanisms and valid mechanistic explanations of human sentence processing. They write:

> "We suggest here an alternative theoretical framework—taking the recent advances in the deep learning of natural language processing seriously, and consider the resulting [ANNs] as plausible models of sentence processing."

(p.8).
Then, they continue by saying that

> "[ANNs] describe syntactic and linguistic processing at a much lower description level compared to common models in psycholinguistics. They could therefore be interpreted as a plausible implementation of the cognitive operations our brain generates during sentence comprehension."

(p.8)
and that "several predictions in humans can be derived from the neural language model, both with respect to behavior and cortical processing" (p. 13). In the same page, the authors take a step further the conclusions of Dong and Toneva (2023), by openly stipulating that

> "the activity of a single unit in the model in response to a feature would map to a large number of spiking neurons in the brain, all responsive to the same feature. Taken together, a single unit in the [ANN] could therefore correspond to possibly more than 106 neurons in the brain."

(p.14).
Moreover, they continue with

> "neural language models are shown to provide precise and testable predictions about both human behavior and its underlying cortical mechanisms, therefore serving as appealing models for both cognitive and mechanistic aspects of human linguistic performance."

(p.14-15).
The clarity and transparency of the article in explicitly expressing what it is usually implicitly given in the studies reviewed so far is much appreciated. By reading these excerpts, it is clear how LLMs are in fact perceived and treated as replicas of not only human cognition but of the human brain itself. The authors openly engage in an explanation for which ANNs are *implementations* of the human brain activity in an artificial device. Simply because perceptrons (ANNs computational units) correlate with some part of the brain activity, they are taken to be an artificial realization of "more than 106 neurons in the brain". Therefore, it is here strikingly clear how the flawed conditional of MP are fully deployed, considering correlating phenomena as identical phenomena - correlation implies identity. Moreover, the principle of multiple realizability is completely disregarded at multiple levels. Not only

underlying mechanisms are understood to be the same for similar behavior (i.e. brain activity responses), but also the implementation level of the phenomenon is considered to be similarly realized, despite ANNs are running on computers' processors made of iron, silicon and gold, while brains are organic materials.

If we consider David Marr's tri-level framework of cognitive capacities (Marr, 2010; Blokpoel and Rooij, 2021) it is possible to observe how lower levels of activity are underdetermined by higher levels, that is, same behavior can be realized by multiple different mechanisms and same mechanisms can be implemented in many different physical systems (see 2.3).
To further explain it, Marr's framework states that a particular cognitive ability can be studied and defined on three different levels: a computational level that specifies *what* type of activity is under investigation what is the input-output relation that constitutes the domain of interest(first level); an algorithmic level, which describes *how* the activity is actually realized in its steps and rules, thus the mechanisms that make the input-output transformations possible (second level); the implementation level, which specifies how the algorithm is realized in a physical system (third level). Each level is underdetermined by its higher one, as each level cannot explain its higher one as multiple realization can carry out the one above.

Marr's tri-levels framework is another useful tool to understand better the fallacies deployed when the principle of multiple realizability is dismissed and the algorithmic and implementation levels are considered as explaining the computational level. This confusion is motivated by the research practice of considering models that fit the data as good explanatory models, thus, considering predictions as explanations. Furthermore, once models that exceptionally fit the data such as LLMs are considered good explanations, it seems only natural to assume that the model is a replication of the target phenomenon (i.e. human linguistic cognition). This is clearly shown in Lakretz et al. (2020).

The same kind of approach is reported also in a cognitive capacity such as language or vision. Goldstein et al. (2020) in the first pages state that their

> "findings provide compelling evidence for shared core computational principles, of prediction and contextual representation, between autoregressive [ANNS] and the human brain, and support a new modeling framework for studying the neural basis of the human language faculty."

(p.3).
They follow by saying that "human predictability scores and GPT2 estima-

tions of predictability were highly correlated. This suggests that GPT2's and humans' next-word predictions are similar in natural contexts." (p.4), adding later on that these correlations "demonstrate how autoregressive [ANNS]' behavior can be used for modeling humans' predictions at behavioral and neural levels" (p. 11).
The problem with this type of approach is the fact that an LLM is treated as a human participant or, at best, as a provider of linguistic material that can be considered human material. That is, LLMs undergo the same analysis that researchers employ for corpus analysis, where mechanistic explanations are hypothesized based on linguistic outputs collected in corpora. What makes the first fallacious and the latter legitimate is the fact that what the LLM produces is an imitation of the surface linguistic behavior of humans, while what researchers look at in corpora is real-world human language. Thus, once again, same surface behavior may be caused by different underlying mechanisms and implementations. This makes generalizations about human language from models' outputs incongruous and inappropriate for a scientific endeavour.

Nevertheless, it is important to point out that, interestingly, Goldstein et al. (2020) conclude by saying:

> "Can [ANNS], such as GPT2, provide insights into the cognitive mechanisms underpinning the human language faculty. We hypothesize that the family of [ANNs]is sharing certain critical computational principles with biological language. This does not imply that they are identical, nor that they share the same circuit architecture. Human brains and [ANNs] share computational principles but they are likely to implement them using radically different neural architectures."

(p.17).
Thus, the authors clearly state that they do not understand the models as replications of human cognition, even though a strong link between the two can be made, as strong as allowing for uncovering the cognitive mechanisms underlying human language. Moreover, they clarify that sharing computational cognitive mechanisms may still show radically different implementation regarding the material and physical structure where the computations are performed. In this regard, they seem to be using Marr's tri-leveled framework to distinguish between different aspects of cognition, since they differentiate the algorithmic level and the implementation level in his terms. Thus, if the reference to Marr is present, they intend to say that whether the implementation happens using electrical circuits made of wires or using organic tissues, it is irrelevant for understanding the human cognitive mechanisms,

provided an understanding of cognition as a series of computations (van Rooij et al., 2023). Therefore, Goldstein et al. (2020) adopt a computational cognitive science perspective, which assumes the possibility of recreating human cognition in computers, given that both systems (human cognition and computers) are information-processing devices (Feigenbaum et al., 1963, as cited in van Rooij et al. (2023)).

Based on this last quote from Goldstein et al. (2020) it is possible to interpret authors' view as considering LLMs as replicas of human cognition at the algorithmic or even computational level. It remains unclear which level they are referring to when affirming that LLMs and the human brain *share computational principles*. However, the replication assumption seems to be operating, otherwise statements such as "deep connections between autoregressive [ANNs] and the human brain" [p.16] would seem irrelevant for a linguistic study.

In the next section (3.4), the relevance for such an inquiry will be proposed, based on the other objective common in many studies here reviewed. Nevertheless, the declared goal of revisiting "classical psycholinguistic models" (Goldstein et al., 2020, p.11) with a LLMs framework to gain new insights into the human language faculty remains a problematic one, as it implies an understanding of LLMs as how-actually models, even though they do not fill in the requirements to be one. Moreover, it keeps open the possibility to assume an identification of the model with the target phenomenon (i.e. language), with a concrete risk of treating the model as a replication of the underlying mechanisms that cause the phenomenon (i.e. human linguistic cognition).

3.4 How human language can help build better LLMs

In reading these studies, one more theme is observable among the objectives of the research. In Galke et al. (2024) and Dong and Toneva (2023), it is explicitly stated how the research goals are bidirectional, as the authors explain how respectively knowledge of language evolution and knowledge about the human brain can contribute significantly to improve the performance of LLMs. Galke et al. (2024) write in their discussion that their

> "findings have further implications for machine learning, where systematic generalization beyond the training distribution is of high interest. Specifically, we show that seeding a learning system with well-structured inputs can improve their ability to sys-

> tematically generalize to new inputs."

(p.11).
Similarly, Dong and Toneva (2023), explicitly state in their conclusion the double intent of their research:

> " We situate our work at the intersection of neuroscience and machine learning, with implications for both fields. **The implications for machine learning**: We show to what extent the models have learned brain-relevant cross-modal connections through the prediction of masked tokens. We provide novel evidence that the cross-modal connections can benefit individual modalities. We identify that current models fall short of capturing multimodal interactions, using the brain as a test bed. We propose a promising and sufficient approach for improvement: fine-tuning a task that requires inference between language and vision."

(p.9.)
In the articles reviewed in 3.2, the theme of improving LLMs remains mostly in the background as the natural starting point of their investigation, as made explicit in Chowdhury and Zamparelli (2018) by affirming that only better performance (i.e. more human-like) will provide the right conditions to ultimately disprove the innatist theory. Thus, understanding the difference existing between LLMs' behavior and humans' behavior in language is certainly primary as no generalizations to human cognitive mechanisms would be possible. In other words, when the relevance of the results of studies seem not to be clear for linguistics, the importance for machine learning is obvious. For instance, when linguistic studies report that LLMs do not show to have the same hierarchical representations that humans have, if they are not adding much to the knowledge of language, they are providing useful insights for the understanding of how LLMs work. For this reason, it seems to be assumed that any findings that report some similarities and differences between the models and humans are useful for data scientists working on improving the models' performance. In Linzen and Leonard (2018), this underlying objective becomes visible by the end of the article when they write: "there are at least two ways in which neural networks can be encouraged to make less errors on syntactically complex sentences."(p.6). The text follows with their proposals to improve the syntactic performance of the model, which appears to be key to argue more confidently against innatism. Therefore, it can be argued that where the linguistic relevance of comparing humans and machines becomes unclear, it is certainly possible to see the valuable

contributions for the field of machine learning.
However, this makes the current analysis more relevant in highlighting how linguistics seem to be more useful to machine learning than the other way around. Where linguists find dubious results regarding human language, the contribution they make to the understanding and development of AI more in general is more obvious, than the contribution to linguistic theory. Indeed, only an underlying - and not always admitted - assumption that these models are replications of human linguistic cognition would identify the real contribution that machine learning can make to linguistics, when using such methodologies.

Nevertheless, it is not intended here that other ways of emplyoing LLMs for linguistic research are equally problematic or redundant as in the cases reviewed here. This study only focuses on a part of linguistic research using LLMs as they would recruit human participants. This is, indeed, the common denominator of all the publications analyzed so far. What is highlighted as unfruitful and arguably harmful for linguistics is this practise of collecting data from LLMs and, by comparing it with human data, drawing conclusions about linguistic cognition that explains human language. This flawed deduction appears in the terminology and logic employed in the papers to be founded on the misconception that LLMs are replicas of human language, thus they can be used to extract direct information about the human language faculty.

3.5 Limitations

It is worth spending some words on some limitations of this study, in order to specify what can be improved and researched further, as well as shortcomings that could be overcome in future investigations.

A first limitation regards the number of articles reviewed, as time-constraints forced the analysis to reduce the number of papers to ensure a thorough analysis of each one of them. However, the limited number of articles reviewed allowed for a deeper understanding of the reasoning behind each study, prioritizing the qualitative aspects of the writing in each paper instead of the quantitative ones. Therefore, one shortcoming of this analysis depends on the limited application of its claim, as it does not have any statistical power to generalize over entire sub-fields of linguistics. Future investigations are needed to assess whether an emerging trend in linguistics is observable.

A second issue the reader may find is the criticism regarding the application of MP and the principle of multiple realizability. Taking the example of the clock, it may be argued that if one cannot safely understand the underlying mechanisms causing the clock to tell the correct time, one may still be interested in studying the time itself, regardless of its internal workings. That is, if a researcher is not interested in language as a cognitive capacity, LLMs may represent a valid tool for its surface behavior. This is the standpoint taken by Grindrod (2024), arguing that linguists may be interested only in the performance, regardless of the competence. He appropriately points out how the dominance of competence in linguistic research, which made language part of cognitive science, is due to Chomsky's framework and its biolinguistics program (see section 3.1). His considerations about language as an organ of which linguists should discover the inner principles of operations automatically marginalized research in language performance, down-grading it as secondary (Grindrod, 2024). On the other hand, socio-linguists might be interested in investigating performance patterns for which LLMs may result as a valid scientific tool.

This thesis does not discard this perspective and the issues highlighted in the articles analyzed are not in contrast with Grindrod (2024) position. His explanation for the dominance of a cognitive perspective in modern linguistics is accurate and it is exactly the perspective that is taken by all the studies here reviewed. Therefore, the main focus in chapter 3 regards specifically cognitive linguistics, here considered involving research programs that investigate linguistic performance to gain insights about competence. In the case of the innatism debate, the question is even more fundamental as laguange acquisition studies are interested in understanding *why* children behave the way they do (i.e. why children learn their first languages so fast and accurately) and *how* they do it. As was pointed out in section 3.1, it might be (rightly) argued that the authors of the studies are mainly concerned with the logical fallacy of explaining innatism on the base of the poverty of the stimulus argument. If LLMs are only used to nourish a feud between research frameworks (Harris, 2021), the relevance of the debate for linguistics becomes questionable. For this reason, it is here assumed that LLMs are considered to be of valuable use to gain insights about language learning.

Finally, the neurolinguistics studies provide good evidence of the replication assumption, which can support language acquisition research in finding scientific explanations in LLMs.

4 Generative AI: a theoretical misuse?

This section discusses the status of LLMs as predictive models used for language generation. It is here argued that generating human-like behavior from models intended to predict that same behavior represents a theoretical misuse. That is, models are used in a way that significantly diverges from their common scientific use, as defined in section 2.4. Moreover, this misuse may be the cause of the above reviewed misinterpretations of what a model can tell us about linguistic cognition.

There is a rather useful book written by a theoretical linguists regarding the status of linguistics among other scientific fields (Scheer, 2011). In his explanation of linguistics and its relations to cognitive science he makes important considerations about the real objectives and driving forces of any scientific endeavour. He writes:

> "The point is that in all adult sciences, the only scientific reality is a representation of the real-world reality, in terms of a drawing and/or in terms of a formula. The relationship between both is often non-trivial, and it typically takes a lot of effort in order to be able to go from one to another: this is what engineering is about. Engineers construct machines on the grounds of a scientific insight."

(Scheer, 2011, p.502)
Further in the chapter he adds:

> "Of course representations such as [the one of an atom] may turn out to be imprecise, incomplete or, according to the degree of incompleteness, simply wrong. But they have been established as the scientific reality, the only reality that counts in science, on grounds that have got nothing to do with any attempt to mimic its real-world properties:"

(p.503)
And this is because in the history of science “progress has always been made when something was understood not when it could be implemented.” (Scheer, 2011, p. 503).
Sheer’s position clearly states the primary role that explanations for real-world phenomena play in the definition of science, paramount to avoid naive empiricist approaches in research, as also pointed out by van Rooij (2022). Thus, in the framework adopted in this thesis, predictive power does not constitute a good scientific model, which has to be first and foremost a tool on our way to understanding.
Nevertheless, without taking into account their scientific utility, models can be also used to prescribe, control and emulate, other than to describe or explain (Blokpoel and Rooij, 2021). For example, models of rational behavior in specific situations, such as what to do in case of an earthquake or rules to follow to prevent the spreading of, are examples of prescriptive models. Furthermore, controlling models may be used by companies to model biases displayed by consumers with the aim of inducing them to buy certain products, and, finally, models can be employed to emulate behaviors, such as the cases reviewed so far in this analysis, where LLMs simulate human-level linguistic production in a conversation via messaging.

By looking at this description, LLMs fall in the category of predictive models that are also emulating models. Their predictive component only involves the mechanism they employ to carry out the task of producing human-like sentences. However, their primary goal is emulation of human language, and if we define our models according to their goals and not their inherent workings, LLMs may be regarded as primarily emulating models. Therefore, once more, nothing explanatory lies underneath their construction. Nevertheless, they are now used in both a part of linguistic research and a part of cognitive science as explanatory models for human (linguistic) cognition, as it was shown in the previous chapters.

This chapter intends to discuss a possible reason for the use of LLMs as explanatory models, even though they do not have the requirements for such a use in linguistics. The aim of this discussion is to go beyond the way linguists use LLMs and understand why linguists chose to to use LLMs in the way it was reported in chapter 3.2. Specifically, the attempt is to provide an explanation that accounts for the theoretical fallacies described in 3.2, by looking at the nature of LLMs, beyond the way they are used. It is indeed common to think that tools are just inert material in our hands, neutral devices that can be used properly or improperly (van der Gun and Guest, 2024). Thus, we often limit ourselves to analyze, as we did in this thesis, what is appropriate and less appropriate behavior with specific tools, given our goals. However, this still serves the idea of the neutrality of technology (van der Gun and Guest, 2024) and does not account for the fact that tools carry with them specific inner properties that shape the way they are used. Their specific properties define the goals that they can be used for and the goals they cannot. They may also shape the way those goals are achieved and the way one designs the steps that will lead to that achievement. For this reason, the property of emulation that LLMs have may be crucial to understand why these models are used the way they are. In fact, their success in emulating language is arguably the reason why logical fallacies are at play in a part of linguistic literature. The emulation is so convincing because they are able to reproduce a real-time conversation with a human, and what makes it possible is the generative aspect of the model. The fact that LLMs are models of language that use a predicting mechanism to generate human-like behavior is here proposed as the core property that induces linguists to consider them how-actually models and replicas of human linguistic cognition.

The fact that LLMs generate language is here regarded as a theoretical misuse. As it was shown in section 2.4 models are approximations of the target phenomenon, a simplification that has to reduce the number of factors at play in the real-world in order to understand a specific fragment of it. Given the intrinsic reductive nature of any type of model, generation of the phenomenon that is modeled is arguably an inappropriate use. Thus, in this case, language models used to generate language should be considered a theoretical misuse that may lead to confusion in scientific applications. The emulating goal of LLMs possible through its generative power may be the property that shapes its use and directs it to a misapplication in linguistics, such as the one that considers it a replication of linguistic cognition.

The problem of language generation is due to the fact that everything that LLMs generate is necessarily a reduction, an approximation, because the model cannot go beyond its simplified nature. This simplification is often perceived as a problem of bias or toxicity and inappropriateness of language. Despite several attempts were made recently to reduce biases, toxic and inappropriate language (Weidinger et al., 2022; Kasirzadeh and Gabriel, 2023; Andrić and Kasirzadeh, 2023), it is often neglected how e.g. bias-free models are in principle impossible to build because of their approximating nature. An approximation may be considered as a bias in itself as models never aim to encompass entire systems with all its nuances with which it appears in reality. For this reason, choices have to be made and in the case of true scientific models, used correctly, the research questions define what is in the model and what can be excluded. For instance, models of language always aim at modeling some aspect of language, such as an internal representation for syntactic structures or the way semantic information and syntactic information interact during sentence parsing. This does not mean that we cannot create bigger models of language that take into account different levels of linguistic and cognitive knowledge, but even in more general models we cannot do without simplifications and reductions. For this reasons, models are not meant to recreate the object they are trying to capture, as they are tools that (should) exist only in relation to our narrations. That is, models are useful only to the extent we can put them in relation to out theories. They help to generate knowledge because they represent a (formal) illustration of that knowledge. And the illustration is not meant to produce an emulation of the target phenomenon to proof its validity. That is, a model of language does not need, nor should learn how to produce language to provide scientific insights about language. This arguably only leads to confusion and the following section explains why.

4.1 Emulating models and the role of perception

Emulation of human activities is not a novelty of our times. Emulation of human activities lies at the core of automating processes that aim at unburden laborers of particularly heavy or repetitive tasks.[1] If machines do not always carry out tasks the way we would do it, the emulation of the output desired is certainly the starting point of any artifact built to substitute (part of) human work or activity. However, the fascination for emulation and reproduction of human activities is generally connected to those ones that are considered typically human and of which we still know very little about. Generally, the types of cognitive activities, broadly described as intelligent behavior, are the hardest to simulate. In line with Scheer (2011), it is probably because we do not have clear scientific insights into what intelligence is and what a cognitive capacity is that engineering it always fails, unless some hidden tricks are used. For instance, the Mechanical Turk (Figure 4.1) is an example of a type of activity generally considered intelligent - playing chess - that was apparently performed by a machine.

The machine seemed to perform a hard cognitive task normally attributed

[1] It is not here discussed the multiple reasons and interests that lead the process of automation in the modern work and society after the first industrial revolution. It is acknowledged how automation has deep common roots with the capitalistic system and its implication in the modern society constitutes a field of study in itself for philosophical, historical and economic research. The reader is referred back to an interesting analysis of automation from a philosophical perspective (Benn and Lazar, 2022).

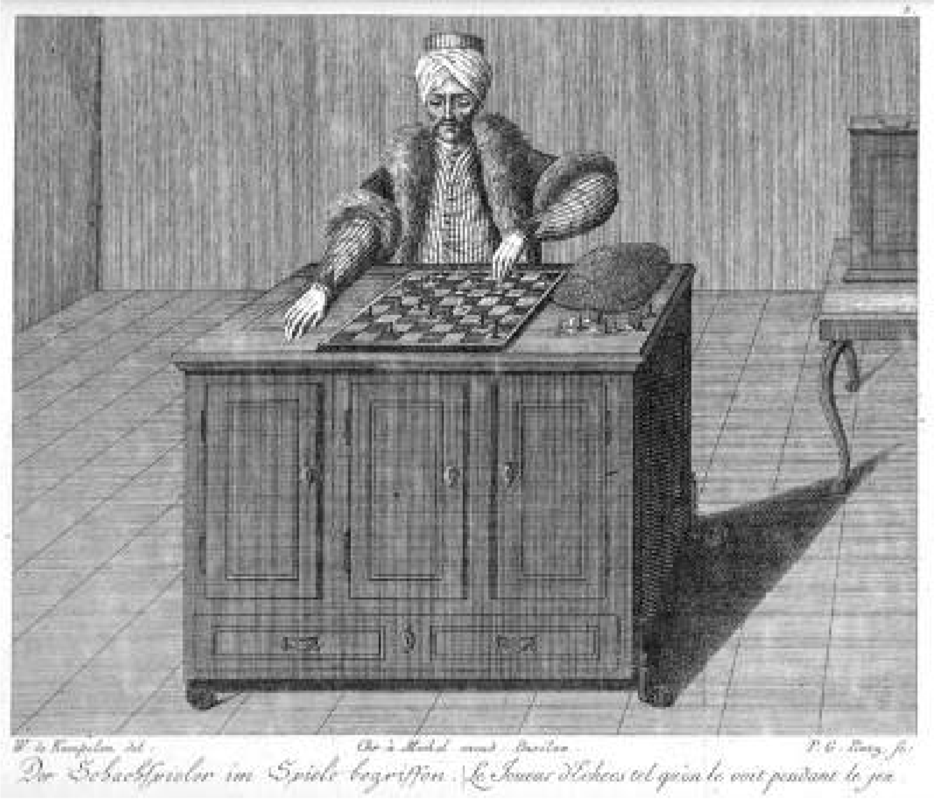
W. de Kempelen del.
Der Schachspieler im Spiele begriffen. Le Joueur d'Echecs tel qu'on le voit pendant le jeu

Figure 4.1: Oriental-looking like chess-player made of wood, also known as the Automaton Chess Player. It was built in 1770 for Maria Teresa of Austria and it was a mechanical illusion as it showed to be able to play chess, competing with humans. After its destruction, it was revealed that the structure of the machine enabled a man to be hidden underneath the chessboard, from where it could see the moves and make the automaton move in response (Patowary, 2022)

to humans, by emulating first the physical realization of that activity (a wooden-made chess player, a human body) and the moves on the board, which were actually controlled by a human player hidden in the box under the chessboard (see Figure 4.1). Thus, the actual *intelligent* behavior was apparently emulated by the automaton, while in reality it was performed by a man. However, to the eyes of the spectators, the machine was emulating human behavior and, by ignoring the existence of the hidden player, they might as well have thought that it was a case of mysterious reduplication of a complex human capacity. Therefore, by not knowing the underlying mechanisms of a surface behavior, it is possible to be led into thinking that simulation must imply duplication, at some level. Thus, by looking at comparable outputs between machines and humans, ignoring the principle of multiple realizability, it is easy to confuse correlation with identity, as it was shown in the previous chapters. This means that specifically the realization of human-like complex behaviors in artificial systems, such as the Turk or an LM, is necessary to lead observers astray. The crucial difference between the Turk and LLMs (other than the fact that the automaton is not a model *per se*) is their application in research, which does not find a parallel in the case of the mechanical chess-player of the XVIII century. In this case, a complex human capacity such as the one of using language is performed well-enough by an algorithm to be mistaken for human production. What is meant here by "complex" is the fact that cognition and its sub-components such as language are not yet understood. That is, cognitive scientists do not agree yet on what cognition is, nor linguists do regarding language (van Rooij et al., 2019; Hauser et al., 2002). What language does and how it does it is still debated, as chapter 3 illustrated for the case of language acquisition. For the case of other cognitive capacity van Rooij et al. (2019) explain that capturing the essence of e.g. decision making without ending up explaining only highly simplified instances of it still results non computable. That is, formalizations that attempt to model decision making results fail to explain real-world decision making processes. However, the fact that our models fail in approximating real-world phenomena is informative enough for scientific progress, because understanding why they fail is a step further in the direction of capturing better what the phenomenon is, what it does and how it does it. Moreover, it should be seen as a fruitful limit that helps us to understand the real usefulness of models, to see them for what they are supposed to be: transitory aids of our own comprehension of the phenomenon, just bridges to an understanding, which is a dynamic, never-ending, recursive process (van Rooij et al., 2023; Guest, 2024).

On the other hand, when this approximating tools, transitory and prone to failure are turned into generating algorithms that recreate the surface behavior of complex human capacities the result may be detrimental for the quality of science. Models that are built to emulate complex cognitive abilities in their surface behavior are bound to fool any kind of observer, because misleading is arguably an implicit goal. That is to say that the perception of an overlapping between the machine performing the task and the human performing the task, in a way that is hard to explain immediately, cannot avoid to be misguided. In the case of the mechanical Turk the deception is intended to entertain, while in the case of LLMs is primarily meant nowadays to make profit. This makes LLMs already unsuitable for scientific research (see 4.2). However, as it was briefly sketched in chapter 2, historically, the evolution of AI is multifaceted and did not linearly develop for only economic interest. It is not here the place to examine the history of AI,[2] but it is worth recalling from chapter 2 that the idea of recreating intelligence in artificial systems, coming from computationalism, played an important role in defining the field of AI as we know it today. Thus, the fact that LLMs represent the technology around which high-tech companies' revenue revolves does not mean that their roots are not fundamentally research-oriented. The profitable results of AI-as-engineering was probably decisive of its success as a branch of computationalism. Generative AI represents the product of AI-as-engineering, exploited for economic interests, and, in its applications to science, it is the core aspect that has to be analyzed to understand where misconceptions may arise.

Recalling here computationalism is meant to highlight how generating human-level language can easily lead linguists into thinking that cognitive aspects of human language were finally replicated in artificial devices. That is to say that there is a theoretical framework that supports the idea of recreating intelligence in artificial systems (Newell and Simon, 1976; van Rooij et al., 2023) and, now, there are artificial systems conversing with humans to the point of convincing them that they have a conscience (Matthews, 2022). Thus, language generation appears to be the proof that the replication idea is indeed possible. Given these observations, it is important to understand whether it is theoretically useful and scientifically appropriate to have models of language that emulate conversation with humans by means of generating language. Put more simply, it is worth considering the appropriateness or at least the necessity of having models generating natural language. If generating language is not necessary to understand it, it may also seem damaging as it fosters the idea that machines can replicate human behavior and, perhaps, even some aspects of its underlying mechanisms. By generating the behavior they are approximating, models risk to become the object of investigation, losing their transitory, disposable status.

Moreover, the level of black-boxing of LLMs is usually quite high and this often makes the causal relation between the underlying mechanisms of the model and its output harder to capture. In this sense, we are not as far from the situation faced by Maria Teresa of Austria when watching the Mechanical Turk in action. A complex human activity such as playing chess is performed by a machine in a strikingly similar way as a human would do and the mechanism can be guessed but not entirely grasped. A similar - even though not completely the same - situation is often faced by linguists who struggle to explain how children learn language and are now witness-

[2] For a complete history of Artificial Intelligence see The Cambridge Handbook of Artificial intelligence (Frankish and Ramsey, 2014)

ing the deployment of chat-bots that learnt how to converse with humans about the linguistic structure of sentences (Beguš et al., 2023). The fact that a model can fluently produce linguistic output with negligible mistakes (at least superficially) is encouraging linguists to revise their theories of language (Baroni, 2022; Piantadosi, 2023). Therefore, it does not seem improper to hypothesize that it is precisely the generative aspect of current LLMs that is at the root of the theoretical fallacies operating in a part of linguistics and CCN. Having models generate human-like output of highly complex and still poorly understood cognitive capacities, such as language, is leading a part of scientists astray, contributing to reinforce an already existing idea that models are end-products of research, instead of dynamic tools for research (van Rooij et al., 2023). With generative AI, models become a research object in itself, with its outputs and mechanisms investigated as natural phenomena.

But if generative AI is not necessary for science and it may be considered to have a detrimental effect on it, why so much effort is spent in building it? To (partially) answer this question, the next section will elaborate further on the nature of LLMs and generative AI as part of a corporate kind of science, led by the techno-economic hegemony of the Big Tech industry.

4.2 LLMs and generative AI as part of a platform society

This chapter discusses the findings of this thesis considering the larger picture of the socio-economic interests in which LLMs are embedded. The research framework of the group led by José van Dijck at Utrecht University will be adopted to explain the deep connections between the developments of AI technologies of the last twenty years and the pervading nature of digital technologies, built and deployed by a handful of multinational American high-tech companies, hereafter also Big Five.

A platform society is defined in van Dijck et al. (2023) and van Dijck (2021) as a digital ecosystem controlled by five American private corporations, widely known under the acronym of GAFAM, G-oogle, A-mazon, F-acebook, A-pple, M-icrosoft. The system behaves like a big platform that survives on the seamless connection between three levels of production and economic control. The companies have growing control over the raw materials needed to build the hardware, the cables and the data centers (the base of the ecosystem). Moreover, they are increasing their presence in the intermediate levels that is constituted by the e.g. cloud infrastructure services, web browsers, search engines, app stores. By completely controlling these two levels of the digital production chain, they also control the access to and the development of sectoral apps and platforms that are used in every sphere of both private and public life (e.g. healthcare, education, entertainment, mobility). This control exercised by the digital infrastructure over every aspect of social, political and economic life is crucial to understand its influence in our society.

One essential property of this ecosystem is that it is data-hungry. Its socio-economic power relies on the continuous collection of data, which flows seamlessly across all layers of its structure - from platforms and applications to search engines and data analytics to data centers, and the other way around -. This highly integrated infrastructure, extensively and intensively entrenched into society has inverted the power relationships between the state and the private sector, making national states and international unions, such as EU, *de facto* increasingly dependent on GAFAM. Academia is not exempted from this dependency.

This situation motivates the choice for referring to this asset of power relation as *platform society.* The extension of its influence across the public and private domains is changing the relations, practices and rules of our societies, making them fit into this digital ecosystem. This is fundamental for the surviving of the infrastructure, since it is based on the monetary exploitation of human data (van Dijck et al., 2023; van Dijck, 2021; Couldry and Mejias, 2019). Human data is merely any traceable human activity, turned into data points, which can be read, interpreted and used by the system (Couldry and Mejias, 2019). For this reason, society should behave like a platform in order to have some monetary value for these companies [3]. This platformization process may be regarded as the source for the sudden growth of data available, on which LLMs are continuously trained.

To understand better the relationship between data collection and LLMs and their embedding in a platform society, it may be useful to return to the introduction of this thesis to recall Turing's forerunning position regarding the progress of computers' capabilities: that the theoretical insights coming from mathematics on which computers are engineered have not changed since Babbage's Analytical Machine and have not changed today with the advent of deep learning. Theoretically, already at Turing's time of writing it was possible to conceive the construction of "an electronic equipment which will 'think for itself', or in which, in biological terms, one could set up a conditional reflex, which would serve as a basis for learning" (Turing, 1950,

[3] For a deeper understanding of the implication of a digitized society, the reader can refer back to the classical book of Shoshana Zuboff, *The Age of Surveillance Capitalism,* published in 2019 by Profile Books.

p.450). Together with important training methods engineered at the beginning of the XXI century (see Rumelhart et al., 1986; Vaswani et al., 2017, for backpropagation and attention in transformers) what enabled these technological innovations to take off toward the type of AI-powered devices and functions easily accessible to a large portion of the global population is the exponential increase of data collection and electricity. As highlighted by Turing himself, electricity has no theoretical value for the invention of the computer, but it drastically changes the possibilities of its practical realizations. Related to electricity is the increasing computing power that modern computers have, visible in its most prominent example from the existence of data centers: entire buildings of computers continuously processing data (Karpathy, 2023). Moreover, as also mentioned by Hinton in one of his recent interviews (CBS Mornings, 2023), it is the availability of immense quantity of human data that makes his teams' learning algorithms so successful (see also Couldry and Mejias, 2019).

The importance of accumulation of human data brings in the necessity to find the causes of this sudden availability. It becomes, thus, clear how van Dijck's description of the platform ecosystem that surround us as individuals and scientists is extremely useful to understand the vital relation that exists between AI and the platform society. Without a continuous collection of data the digital infrastructure could not survive for long and AI systems like LLMs capable of engaging in conversations with humans (e.g. Chat-GPT) could not be conceivable. Data is ensured by reshaping society as a platform, which ensures that human activities can be interpreted and understood by the digital infrastructure. Thus, data is provided by the users that increase in quantity and typology, as increasingly diversified sectors of society are being absorbed by the digital ecosystem, massively migrating their institutional infrastructures (e.g. medical centers, postal services, educational environments) to the digital one created by external parties - external as in extra-governmental and, for the majority of the nations, foreign, as the ecosystem is American. This "migration online", as it is often called in ordinary terms, increases the data availability that flows unconstrained across all the layers of the ecosystem. Importantly, what is less obvious from this visualization is the joint action of the Big Five, as they keep a collaborative competition type of relationship, for which they compete for the highest economic power, maintaining a collaborative approach as far as data collection is involved. van Dijck (2021) explains that: "the Big Five are also *interdependent* (...) Interdependencies turn the Big Five platforms into "coordinating competitors"— a form of "coopetition" that easily escapes scrutiny by regulatory agencies who tend to focus on individual firms (Daidj and Egert, 2018; Kostis, 2018)" (p. 2089).

This socio-technical monopoly is, thus, the key aspect that drives the current AI development and it arguably affects autonomy of academic research (Giulimondi, 2024; Kerssens and van Dijck, 2022; Ahmed et al., 2023). The embedding of LLMs and its applications in an industry-driven framework, which is product-oriented and profit-oriented, should not be considered out of the domain of meta-theoretical research. When considering the theoretical choices that motivate our research, the reasons for the choice of a specific methodology or the theoretical framework we choose to justify our interpretation of data and our conclusions, it is not of secondary importance to gain awareness of the infrastructure that made our study possible. It is important to reflect on the implications of letting science be dependent on private corporations that pursue very different objectives than academic research. It is clear from what was outlined in this chapter that the dependency concerns the tools and not simply a financial support. And the tools were shown to be a determining aspect of the outcome of linguistic research, in this analysis. More specifically, the generative nature of LLMs, together with their architecture and deep learning methods developed regardless of any linguistic knowledge, represents a feature proper of an industrial framework, whose models are developed for the purpose of selling and not understanding phenomena. If those models may have a second life in helping scientific investigations, it is arguably a matter of coincidence that will not fail to miss its chance of profit. In this sense, specific private companies that already have such a big influence on our society can shape also academic research through the tools they deploy and universities eventually adopt. If tools are not neutral and their inherent properties direct the use we make of them, we are indirectly guided by the mindset of their builders. That is, if builders embrace AI-as-engineering as their theoretical framework and their products reflect this approach, any research using these tools is bound to nurture and reinforce this framework.

In this sense, contemporary AI tools, as the diamond tip of modern technology, represent a research framework in itself, with its own hypothesis and methods. Generative AI uses approximations of human data to make machines perform language, even though linguists do not yet agree on what this specific cognitive phenomenon is (van Rooij et al., 2023; Rich et al., 2021). Language generation can be, thus, considered as part of a “corporate science” research framework (Katzir, 2023), arguably influencing the scientific understanding of language.

In conclusion, language generation from an approximation of a model is here considered to serve the interests of the industry and indirectly affecting academic research. As described in chapter 4, generative AI may be the reason why linguists are willing to revise long-standing, scientifically sound theories of language because an artificial system is able to mimic human language using very different mechanisms from the one that were assumed by linguists (Baroni, 2022; Piantadosi, 2023). Besides the debate around innatism, a part of linguistic research is inclined to embrace the replication assumption simply because a model can generate human-like behavior. This generative aspect is part of what it may be referred to as a "BigTech framework" and it is wise to consider how this is not neutral, as technology - unlike most common opinions - is never entirely neutral (van der Gun and Guest, 2024). The non-neutrality of generative AI, thus, of LLMs and their down-scaled versions (see, for example, the here reviewed BabyBERTa) may negatively affect the objectivity of academic research. It is, thus, important to reflect on whether linguistic research may be also exposed in some of its programs to this kind of enframing and develop research frameworks that enable linguists to think more in depth about the influence of industry in their own research.

5 Some final remarks and conclusions

This thesis engaged in a thorough analysis of ten linguistic articles investigating the neural and cognitive aspects of language. A replication assumption was observed to be present in most of the studies reviewed and this was argued to be caused by the generative aspect of LLMs, producing human-level linguistic output to converse with humans. This generative property of current LLMs was put in relation to the digital ecosystem (platform society) in which they are embedded and the fundamental characteristics that they share with it, arguing for a diminished autonomy of academic research when LLMs are employed for scientific studies.

The discussion of the relation between the platform society and AI and the implications for academic research is not intended to discredit the efforts and scientific outcomes of current research employing LLMs and ANNs more in general. The intent of the last chapter is highlighting the importance of meta-theoretical thinking also in light of the socio-economic dynamics guiding our society. It is often common to witness a tendency in academic environments to dismiss these considerations as outside of the scope of science and inappropriate for scientific debate. On the contrary, socio-political matters are often the starting engine of most technological development and the implication for the directions of science are fundamental to frame our research efforts in the bigger picture of societal impact and change. Moreover, academia plays a central role in the advancement of scientific knowledge at the service of the society, but it is also embedded in that same society and it would be naive to think that it is immune to the pressures and demands of other social actors. For this reason, the last part of this thesis constitutes an effort to understand linguistics as part of a bigger scientific development today mostly propelled by private corporations (Ahmed et al., 2023). The meta-theoretical framework of this analysis should not miss the reasons and premises that may be at the core of many theoretical fallacies.

In conclusion, it was reported how a part of linguistics manifests an inclination to embrace the replication assumption, here defined as considering LLMs a replication of human linguistic cognition, a recreation of it in an artificial system. This replication assumption has its roots in a particular evolution of computationalism, now referred to as AI-as-engineering, the idea that human cognition can be recreated in artificial systems. Ten articles were reviewed in which this assumption was found in different forms and different degrees of explicitness. Overall, LLMs are minimally considered how-actually models and maximally pure replications of linguistic cognition. Finally, this theoretical fallacy was put in relation with the generative aspect of LLMs, an aspect considered a theoretical misuse. The misuse was explained as the failure to recognize the scientific status of models as simple aids for scientists' understanding of the target phenomena (i.e. language). Scientific models are approximations of phenomena, meant to simplify and reduce reality for the purpose of understanding some aspects of its workings. For this reason, generating complex human behaviors from these reductions is considered a scientific misuse that is arguably misleading linguists in considering LLMs an instantiation of human language that can be compared to humans to gain some insights about their cognition. Finally, generative AI is here considered to be part of a BigTech framework, where the interests of the industry are influencing the outcome of academic research through the adoption at scale of industry-based tools, such as LLMs.

The meta-theoretical framework adopted in this thesis, thus, considers the additional dimension of the research environment in which LLMs are designed, built and deployed, highlighting the influence it can have on the autonomous development of academic research, meant not to serve the economic interests of a few private companies but the social well-being.

Bibliography

Ahmed, N., Wahed, M., and Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, 379(6635):884–886.

Andrić, K. and Kasirzadeh, A. (2023). Reconciling Governmental Use of Online Targeting With Democracy. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1871–1881, New York, NY, USA. Association for Computing Machinery.

Aten, J. J. (2024). 'Scientists should be careful when interpreting results of AI models'. `https://www.staff.universiteitleiden.nl/news/2024/04/scientists-should-be-careful-when-interpreting-results-of-ai-models`. [Comment on the webpage of Leiden University News].

Auguera y Arcas, B. (2022, September 2). Artificial neural networks are making strides towards consciousness, according to Blaise Agüera y Arcas. *The Economist*.

Backus, A., Cohen, M., Cohn, N., Faber, M., Krahmer, E., Laparle, S., Maier, E., Miltenburg, E. v., Roelofsen, F., Sciubba, E., Scholman, M., Shterionov, D., Sie, M., Tomas, F., Vanmassenhove, E., Venhuizen, N., and Vos, C. d. (2023). Minds: Big questions for linguistics in the age of AI. *Linguistics in the Netherlands*, 40(1):301–308. Publisher: John Benjamins.

Baroni, M. (2022). On the Proper Role of Linguistically Oriented Deep Net Analysis in Linguistic Theorising. In Lappin, S. and Bernardy, J. P., editors, *Algebraic Structures in Natural Language*, pages 1–16. Taylor and Francis.

Beguš, G., Dąbkowski, M., and Rhodes, R. (2023). Large Linguistic Models: Analyzing theoretical linguistic abilities of LLMs. `https://arxiv.org/abs/2305.00948`.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Benn, C. and Lazar, S. (2022). What's Wrong with Automated Influence. *Canadian Journal of Philosophy*, 52(1):125–148.

Blokpoel, M. and Rooij, I. J. E. I. v. (2021). Theoretical modeling for cognitive science and psychology. *NIAS - The Netherlands Institute for Advanced Study in the Humanities and Social Sciences.*

Butcher, C. and Goldin-Meadow, S. (2000). *Gesture and the transition from one- to two-word speech: when hand and mouth come together*, page 235–258. Language Culture and Cognition. Cambridge University Press.

Börner, K., Boyack, K. W., Milojević, S., and Morris, S. (2012). An Introduction to Modeling Science: Basic Model Types, Key Definitions, and a General Framework for the Comparison of Process Models. In Scharnhorst, A., Börner, K., and van den Besselaar, P., editors, *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*, pages 3–22. Springer, Berlin, Heidelberg.

CBS Mornings (2023). Full interview: "Godfather of artificial intelligence" talks impact and potential of AI. `https://www.youtube.com/watch?v=qpoRO378qRY`.

Chomsky, N. (1988). Knowledge of language: its nature, origin, and use. *The Philosophical Review*, 97:567.

Chowdhury, S. A. and Zamparelli, R. (2018). Rnn simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th international conference on computational linguistics*, pages 133–144.

Clark, A. and Sakas, W. G. (2010). Computational Models of First Language Acquisition Special Issue of Research on Language and Computation. *Research on Language and Computation*, 8(2):101–106.

Clark, H. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Couldry, N. and Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, 20(4):336–349.

Crain, S. and Nakayama, M. J. (1987). Structure dependence in grammar formation. *Language*, 63:522–543.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dong, D. T. and Toneva, M. (2023). Vision-Language Integration in Multimodal Video Transformers (Partially) Aligns with the Brain. arXiv:2311.07766 [IN PRESS].

Fabbro, F. (2004). *Neuropedagogia delle lingue.* Astrolabio-Ubaldini.

Feigenbaum, E. A., Feldman, J., et al. (1963). *Computers and thought*, volume 37. New York McGraw-Hill.

Ferrari, F., van Dijck, J., and van den Bosch, A. (2023). Foundation models and the privatization of public knowledge. *Nature Machine Intelligence*, 5(8):818–820.

Frankish, K. and Ramsey, W. M. (2014). *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press.

Galke, L., Ram, Y., and Raviv, L. (2024). What Makes a Language Easy to Deep-Learn? arXiv:2302.12239 [IN PRESS].

Galke, L. and Raviv, L. (2024). Emergent communication and learning pressures in language models: a language evolution perspective. arXiv:2403.14427 [IN PRESS].

Giulimondi, A. (2024). Why academia should cut back general enthusiasm about CAs. In Hosseini-Kivanani, N., Höhn, S., Anastasiou, D., Migge, B., Soltan, A., Dippold, D., Kamlovskaya, E., and Philippy, F., editors, *Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024)*, pages 9–15, St Julians, Malta. Association for Computational Linguistics.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, 10(5):447–474.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2020). Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *BioRxiv*, pages 2020–12.

Grindrod, J. (2024). Modelling language. *arXiv preprint arXiv:2404.09579*.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without broca's area. *Behavioral and Brain Sciences*, 23:1 – 21.

Guasti, M. T. (2002). *Language Acquisition: The Growth of Grammar*. MIT Press, Cambridge, United States.

Guest, O. (2024). What makes a good theory, and how do we make a theory good? *Computational Brain & Behavior*, pages 1–15.

Guest, O. and Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 16(4):789–802.

Guest, O. and Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*, 6(2):213–227.

Harari, Y. N. (2023). Yuval Noah Harari argues that AI has hacked the operating system of human civilisation. *The Economist*.

Harris, R. A. (2021). *The linguistics wars: Chomsky, Lakoff, and the battle over deep structure*. Oxford University Press.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579.

Holle, H. and Gunter, T. C. (2007). The Role of Iconic Gestures in Speech Disambiguation: ERP Evidence. *Journal of Cognitive Neuroscience*, 19(7):1175–1192.

Holler, J. (2022). Visual bodily signals as core devices for coordinating minds in interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377.

Holler, J. and Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, 2003(146).

Hostetter, A. B. and Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3):495–514.

Huebner, P. A., Sulem, E., Cynthia, F., and Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In Bisazza, A. and Abend, O., editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Karpathy, A. (2023). [1hr Talk] Intro to Large Language Models [Video]. `https://www.youtube.com/watch?v=zjkBMFhNj_g`.

Kasirzadeh, A. and Gabriel, I. (2023). In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, 36(2):27.

Katzir, R. (2023). Why Large Language Models Are Poor Theories of Human Linguistic Cognition: A Reply to Piantadosi. *Biolinguistics*, 17:1–12.

Kerssens, N. and van Dijck, J. (2022). Governed by Edtech? Valuing Pedagogical Autonomy in a Platform Society. *Harvard Educational Review*, 92(2):284–303.

Kirby, S. (2002). *Learning, bottlenecks and the evolution of recursive syntax*, page 173–204. Cambridge University Press.

Kodner, J., Payne, S., and Heinz, J. (2023). Why Linguistics Will Thrive in the 21st Century: A Reply to Piantadosi (2023). `http://arxiv.org/abs/2308.03228`.

Lakretz, Y., Dehaene, S., and King, J.-R. (2020). What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, 22(4).

Linzen, T. and Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Linzen, T. and Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. *Annual Meeting of the Cognitive Science Society*.

Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., Zhang, Y., Wang, J., Gao, X., Zhong, T., et al. (2024). Understanding llms: A comprehensive overview from training to inference. *arXiv preprint arXiv:2401.02038*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Logan, J. A. R., Justice, L. M., Yumuş, M., and Chaparro-Moreno, L. J. (2019). When Children Are Not Read to at Home: The Million Word Gap. *Journal of Developmental & Behavioral Pediatrics*, 40(5):383.

Marcus, G. (2018). Deep Learning: A Critical Appraisal. arXiv:1801.00631 [cs, stat].

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information.* MIT press.

Martínez, H. J. V., Heuser, A. L., Yang, C., and Kodner, J. (2023). Evaluating Neural Language Models as Cognitive Models of Language Acquisition. *GENBENCH*.

Matthews, D. (2022). Does this AI know it's alive? *Vox*.

McCoy, R. T., Frank, R., and Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 40.

McNeill, D. (1992). *Hand and Mind: What gestures reveal about thought.* The University of Chicago Press.

Nature (2023). Language models and linguistic theories beyond words. *Nature Machine Intelligence*, 5:677 – 678.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2024). A Comprehensive Overview of Large Language Models. arXiv:2307.06435 [cs].

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, 19(3):113–126.

Oudeyer, P.-Y., Kachergis, G., and Schueller, W. (2019). Computational and robotic models of early language development: a review. *International Handbook of Language Acquisition*, pages 76–101.

Patowary, K. (2022). The Mechanical Turk: An 18th Century Chess Playing Robot. *Amusing Planet*.

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, 9:1109.

Perniss, P., Thompson, R. L., and Vigliocco, G. (2010). Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages. *Frontiers in Psychology*, 1.

Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *LingBuzz Preprint, lingbuzz*, 7180.

Rich, P., de Haan, R., Wareham, T., and van Rooij, I. (2021). How hard is cognitive science? In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Scheer, T. (2011). *A Guide to Morphosyntax-Phonology Interface Theories.* De Gruyter Mouton, Berlin, New York.

Schlinger, H. D. (2009). Theory of Mind: An Overview and Behavioral Perspective. *The Psychological Record*, 59(3):435–448.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., and Young, M. (2014). Machine learning: The high interest credit card of technical debt. In *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop).*

Seals, S. and Shalin, V. L. (2023a). Discourse over discourse: The need for an expanded pragmatic focus in conversational ai. *arXiv preprint arXiv:2304.14543.*

Seals, S. M. and Shalin, V. L. (2023b). Expanding the Set of Pragmatic Considerations in Conversational AI. arXiv:2310.18435 version: 1.

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., and van Gerven, M. a. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785.

Slonimska, A., Ozyurek, A., and Campisi, E. (2015). Ostensive signals: Markers of communicative relevance of gesture during demonstration to adults and children. In *4th GESPIN-Gesture & Speech in Interaction Conference*, pages 217–222. Universite of Nantes.

Slonimska, A., Özyürek, A., and Capirci, O. (2020). The role of iconicity and simultaneity for efficient communication: The case of Italian Sign Language (LIS). *Cognition*, 200:104246.

Slonimska, A., Özyürek, A., and Capirci, O. (2022). Simultaneity as an Emergent Property of Efficient Communication in Language: A Comparison of Silent Gesture and Sign Language. *Cognitive Science*, 46(5):e13133.

Snow, C. E. and Ferguson, C. A., editors (1977). *Talking to Children: Language Input and Acquisition*. Cambridge University Press.

Sullivan, E. (2022). Understanding From Machine Learning Models. *British Journal for the Philosophy of Science*, 73(1):109–133. University of Chicago Press.

Tagliaferri, C. C., Paperno, D., Gatt, A., and Axioti, S. (2023). The scenario refiner: Grounding subjects in images at the morphological level. In Aggarwal, P., Ala\ccam, \., Silberer, C., Zarrie\ss, S., and Zesch, T., editors, *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 25–33, Ingolstadt, Germany. Association for Computational Lingustics.

Tomasello, M. (2009). The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.

Treccani (2003). Modello. `https://www.treccani.it/vocabolario/modello/,https://www.treccani.it/vocabolario/modello/`.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):433–460.

Valian, V. (2009). Innateness and learnability. *Handbook of child language*, pages 15–34.

van der Gun, L. and Guest, O. (2024). Artificial intelligence: Panacea or non-intentional dehumanisation? *Journal of Human-Technology Relations*, 2.

van Dijck, J. (2021). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society*, 23(9):2801–2819.

van Dijck, J., de Winkel, T., and Schäfer, M. T. (2023). Deplatformization and the governance of the platform ecosystem. *New Media & Society*, 25(12):3438–3454.

van Dijk, B. M. A., Kouwenhoven, T., Spruit, M. R., and van Duijn, M. J. (2023). Large Language Models: The Need for Nuance in Current Debates and a Pragmatic Perspective on Understanding. arXiv.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1(3):127–128. Nature Publishing Group.

van Rooij, I., Blokpoel, M., Kwisthout, J., and Wareham, T. (2019). *Introduction*, page 3–22. Cambridge University Press.

van Rooij, I., Guest, O., Adolfi, F. G., Haan, R. d., Kolokolova, A., and Rich, P. (2023). Reclaiming AI as a theoretical tool for cognitive science. Publisher: OSF.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Volterra, V., Caselli, M. C., Capirci, O., and Pizzuto, E. (2004). Gesture and the emergence and development of language. *Beyond nature-nurture*, pages 53–90.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Warstadt, A. and Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., and Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229, Seoul Republic of Korea. ACM.

Wertheimer, T. (2022). Blake Lemoine: Google fires engineer who said AI tech has feelings. *BBC News*.

White, L. (2020). Linguistic theory, universal grammar, and second language acquisition. In *Theories in second language acquisition*, pages 19–39. Routledge.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451–456. Publisher: Elsevier.

Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., and Choi, Y. (2022). Merlot reserve: Neural script knowledge through vision and language and sound. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16354–16366.

A : Keywords

This is a list of keywords that guided the analysis in the selection of relevant statements in the papers reviewed. The keywords are linguistic cues that signal in the text the presence of underlying assumptions guiding the research.

- cognitive plausibility
- plausible models
- developmentally plausible
- plausible implementations
- mechanistic explanation
- models of cognitive and mechanistic aspects
- cognitive mechanisms
- underlying mechanisms
- competence
- replicate
- possibly akin
- brain-relevant
- positive results
- proof of concepts
- generalize to humans

- useful evidence (of human learning)
- similar to humans

B : Categories

This is the list of the data categories collected for the analysis. For the complete analysis you can visit this link.

- Authors
- Year of publication
- DOI
- Number of citations
- Research subfield
- (L)LM used.
- Research questions
- Methodology
- Relevant ciatations and keywords
- Critical summary of the use of (L)LMs

C : Articles reviewed

This is the complete list of articles reviewed.

- Warstadt, A. and Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In Algebraic structures in natural language, pages 17–60. CRC Press.

- McCoy, R. T., Frank, R., and Linzen, T. (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks. In Proceedings of the Annual Meeting of the Cognitive Science Society, volume 40.

- Linzen, T. and Leonard, B. (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. Annual Meeting of the Cognitive Science Society.

- Chowdhury, S. A. and Zamparelli, R. (2018). Rnn simulations of grammati- cality judgments on long-distance dependencies. In Proceedings of the 27th international conference on computational linguistics, pages 133–144.

- Huebner, P. A., Sulem, E., Cynthia, F., and Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In Bisazza, A. and Abend, O., editors, Proceedings of the 25th Conference on Computational Natural Language Learning, pages 624–646, Online. As- sociation for Computational Linguistics.

- Galke, L., Ram, Y., and Raviv, L. (2024). What Makes a Language Easy to Deep-Learn? arXiv:2302.12239 [IN PRESS].

- Galke, L. and Raviv, L. (2024). Emergent communication and learning pressures in language models: a language evolution perspective. arXiv:2403.14427 [IN PRESS].

- Lakretz, Y., Dehaene, S., and King, J.-R. (2020). What limits our capac- ity to process nested long-range dependencies in sentence comprehension? Entropy, 22(4).

- Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2020). Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. BioRxiv, pages 2020–12.

- Dong, D. T. and Toneva, M. (2023). Vision-Language Integration in Multimodal Video Transformers (Partially) Aligns with the Brain. arXiv:2311.07766 [IN PRESS].