philosophies

MDPI

Article

The Simulative Role of Neural Language Models in Brain Language Processing

Nicola Angius *, Pietro Perconti, Alessio Plebe and Alessandro Acciai

Department of Cognitive Science, University of Messina, Via Concezione 8, 98121 Messina, Italy; pietro.perconti@unime.it (P.P.); alessio.plebe@unime.it (A.P.); alessandro.acciai@studenti.unime.it (A.A.)
* Correspondence: nicola.angius@unime.it

Abstract: This paper provides an epistemological and methodological analysis of the recent practice of using neural language models to simulate brain language processing. It is argued that, on the one hand, this practice can be understood as an instance of the traditional simulative method in artificial intelligence, following a mechanistic understanding of the mind; on the other hand, that it modifies the simulative method significantly. Firstly, neural language models are introduced; a study case showing how neural language models are being applied in cognitive neuroscience for simulative purposes is then presented; after recalling the main epistemological features of the simulative method in artificial intelligence, it is finally highlighted how the epistemic opacity of neural language models is tackled by using the brain itself to simulate the neural language model and to test hypotheses about it, in what is called here a co-simulation.

Keywords: simulative artificial intelligence; synthetic method; mechanism; neural language models; brain language processing; deep learning

check for updates

Citation: Angius, N.; Perconti, P.; Plebe, A.; Acciai, A. The Simulative Role of Neural Language Models in Brain Language Processing. *Philosophies* **2024**, *9*, 137. https://doi.org/10.3390/philosophies9050137

Academic Editors: Marcin J. Schroeder and Gordana Dodig-Crnkovic

Received: 10 July 2024
Revised: 23 August 2024
Accepted: 28 August 2024
Published: 29 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The use of machines to predict and explain the intelligent and adaptive behaviours of biological systems traces back to the birth, in the middle of the twentieth century, of cybernetics, due to the groundbreaking work of Norbert Wiener [1]. Cybernetics was also conceived as an attempt to promote a *mechanistic* view of living systems in apparent contrast with the vitalism of Henri Bergson and the use of the "vital force" principle to explain natural evolution and adaptation [2]. The epistemological setting of cybernetics has been fully inherited by Artificial Intelligence (AI), especially in the simulative approach of the pioneers Hallen Newell and Herbert Symon. The so-called simulative, or *synthetic*, method in AI amounts to using computational systems to test cognitive hypotheses about some natural cognitive system [3]. The synthetic method influenced research in AI, under both the symbolic and sub-symbolic paradigm, and in robotics.[1]

AI is now living what has been a called a *Renaissance* era [4], thanks to the unexpected success of *Deep Learning* (DL). Roughly speaking, two main paths can be identified along which the resurgence of AI has unfolded in the last ten years. In the first five years, the most successful path was vision, leading for the first time to artificial systems with a visual recognition ability similar to that of humans [5–9], arousing surprise and interest in the science of vision [10–12]. Five years later, it was the turn of language, a path opened by the Transformer model [13], quickly followed by various evolutions and variants [14–17], generically called here Neural Language Models (NLMs). In this case too, the sudden and unexpected availability of artificial systems with linguistic performances not so far from human ones has deeply shaken the scientific community of language scholars [18–22].

The success of DL in crucial cognitive tasks such as vision and language has prompted different reactions from the cognitive neuroscience community, ranging from acknowledgment [11], to curiosity [12], to refusal [23]. One main reason for such different attitudes towards DL is that whereas traditional Artificial Neural Networks (ANNs) were explicitly inspired by the functioning of the brain, the development of the Transformer architecture has not been influenced by the functional or structural organization of the brain. And nonetheless, a new line of research in cognitive neuroscience uses Transformer-based models to simulate brain activities. More specifically, NLMs are being used to predict cortex activations while processing language [24–26].

This paper intends to show how the application of DL networks in the study of brain language processing can be understood, from an epistemological and methodological point of view, as an instance of the simulative method as considered in [3], in continuity with the mechanistic approaches in the philosophy of cognitive science. In particular, it is examined how NLMs are used to simulate human agents involved in linguistic tasks, providing predictions about the human cognitive system.

The main aim of this paper is, nonetheless, highlighting significant methodological differences that arise when DL is involved in simulative tasks. In traditional simulative AI, cognitive hypotheses are tested by experimenting on the simulative system, as long as one cannot directly experiment on the simulated system, due to ethical concerns or when the simulated system is *epistemically opaque*. However, epistemic opacity and non-interpretability is one essential feature of DL models as well [27]; this marks a significant difference between NLMs and the simulative programs of symbolic AI or the ANN of the connectionist approach. It is argued here that, in order to overcome the limited intepretability of NLMs when used to simulate brain language processing, the brain itself is used as a model of the NLM in what is called here a *co-simulation*. The idea of using a natural cognitive system to simulate an artificial computational one strengthens even more the mechanistic view of the human mind.

This paper is organized as follows. Section 2 introduces NLMs and the Transformer architecture; Section 3 shows how NLMs are being used in cognitive neuroscience for simulative purposes in the context of brain language processing; Section 4 underlines the main epistemological features of the simulative method in AI and bio-robotics; Section 5 analyses how the simulative method is applied and modified in NLM simulations; finally, Section 6 concludes the paper.

2. Neural Language Models

The conquest of natural language has been one of the most difficult challenges for AI, and for a long time, ANNs have played a secondary role compared to conventional Natural Language Processing. The first attempt to integrate ANNs into natural language processing was undertaken by [28], concentrating on inflectional morphology. Their aim was to show, through an artificial model, that learning the morphology of the past tense of English verbs does not necessitate explicit or innate rules, but it is instead acquired from experience. Their model succeeded and was able to replicate the typical learning curves observed in young children. However, Rumelhart and McClelland faced a significant challenge in employing ANNs for language processing due to a seemingly irreconcilable discrepancy between the two formats. Language is an ordered sequence of auditory signals (in the case of spoken language) or symbols (in the case of written language), whereas a neural layer is a real vector with a fixed dimension. This creates a problem in encoding an arbitrary length datum (the word) with a fixed-dimension vector (the neural layer), even for models restricted to the processing of single words.

A second challenge in applying ANNs to natural language processing is that representing words with neural vectors becomes more problematic when moving from single-word morphology to syntax. Feedforward ANNs are static, making it difficult to establish a sense of order for multiple words in a sentence.

An additional challenge for traditional ANNs arises from the very technique that determined their success in the ’90s: backpropagation learning [29]. Efficient backpropagation requires tasks where inputs and outputs are clearly identifiable, and examples of these input-output pairs must be available, i.e., supervised training. However, the ability to understand language, and even more so to produce it, extends beyond tasks where the necessary inputs and outputs for supervised training can be distinctly identified.

Fueling the confidence in those who, despite these negative premises, have persevered, is the fact that the symbolic nature of language seems antithetical even to the neurons of our brain, which apparently have solved these problems very well. This confidence was well placed, and finally crowned by the Transformer architecture [13] combining several effective strategies to cope with the symbolic nature of natural language. The first strategy is *word embedding*, which learns from examples to optimally convert words into vectors of neural activity. Introduced by [30], its key feature is that the vector representation is semantically meaningful. These numerical vectors can be manipulated in ways that respect lexical semantics. For instance, let vector $\vec{w}(\cdot)$ represent the word embedding transformation and let $\vec{w}(\texttt{king})$ be the vector for the word 'king'; by subtracting from it the vector $\vec{w}(\texttt{male})$ for 'male' and adding the vector $\vec{w}(\texttt{female})$ for 'female', one obtains vector $\vec{q}$:

$$\vec{q} = \vec{w}(\texttt{king}) - \vec{w}(\texttt{male}) + \vec{w}(\texttt{female})$$

which is closer to the vector $\vec{w}(\texttt{queen})$ for the word 'queen' than to any other word embedding vector.

The second strategy is the *attention* mechanism, firstly introduced by [31] in the framework of patter recognition and later on, in the context of language generation, by [13]. This method dynamically identifies relevant information and relationships among words in a sentence. The Transformer employs these strategies in an innovative way. Firstly, word embedding is learned as the entire neural model processes corpora. Secondly, the attention mechanism completely replaces recursion, allowing all words, along with their vector embeddings, to be simultaneously presented as input.

Furthermore, the Transformer incorporates an elegant solution to bypass supervised learning, as introduced by [32]: the concept of the *autoencoder*. This deceptively simple idea involves assigning the ANN the task of reproducing its own input as output. The architecture implementing this concept is typically organized into two components. The encoder generates an internal representation of the input, while the decoder reproduces the output from this representation, which coincides with the input. The popular term "stochastic parrots" [33] for Transformer models originates from this autoencoder structure. Although the term accurately reflects the training technique, it becomes irrelevant when used derogatorily towards NLMs. This exemplifies what [34] have termed a *Redescription Fallacy*, where a NLM's skills and abilities are judged based on irrelevant characteristics, such as the training strategy in this case.

The remarkable efficiency of the Transformer has led to many variations, including ViT *Vision Transformer* [35] and BERT (*Bidirectional Encoder Representations from Transformers*), where attention is applied to both the left and right side of the current word [14]. The original Transformer was designed for translation, so it includes an encoder for the input text and a decoder for the text generated in a different language. A simplification was later adopted by GPT (Generative Pre-trained Transformer), which consists only of a decoder part, primarily for generating text by completing a given prompt [15]. The popular public interface ChatGPT is based on later models of the GPT family [16]. The autoencoding strategy during learning is the task of just predicting the next token in a text. In a strictly mathematical sense, the output of the Transformer is the probability of tokens being generated at the next time step. It is important to note that often this interpretation—although entirely correct in itself—is mistakenly regarded as the overall task performed by the Transformer, thus leading to a misleading underestimation of it. Similarly, it would not be incorrect to assert that when a person writes a word, it corresponds to the highest probability in a space of brain neural activations of the entire vocabulary. But if one were to limit oneself to this to account, for example, for the words we authors put one after the other in this sentence, it would be a truly disappointing explanation.

The subsequent description pertains to the streamlined GPT architecture, with an overall scheme shown in Figure 1. The input text consists of *tokens*t_i, where each token is an integer index into the vocabulary, which comprises words, punctuation marks, and parts of words. The vocabulary size N typically includes several tens of thousands of entries. A crucial operation on the input token is embedding, performed with the embedding matrix $\mathrm{W}_E \in \mathbb{R}^{D\times N}$, where D is the embedding dimension. For a token t_i in the input stream, the embedded vector is computed as follows:

$$\vec{x}_i = \mathrm{W}_E^{(t_i)} + p(i) \tag{1}$$

where $\mathrm{W}_E^{(j)}$ is the j-th column of W_E and $p(\cdot) : \mathbb{N} \to \mathbb{R}^d$ is a function that encodes the position of the token inside the stream of text.

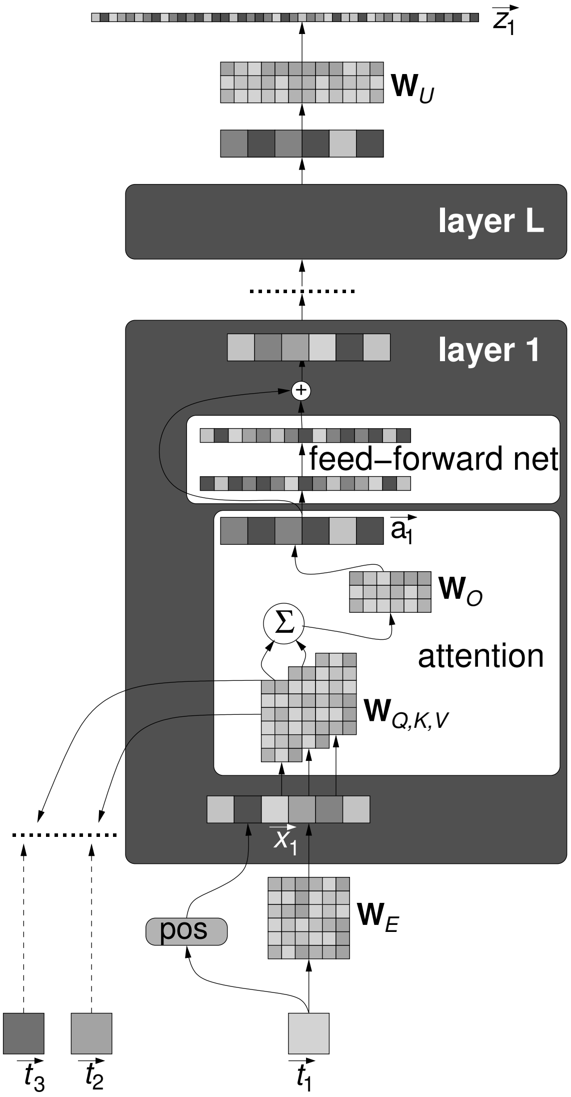

Figure 1. A simplified scheme of the overall Transformer architecture. All components are described in the text.

The model consists of a chain of L layers, with each layer comprising an attention block followed by a feedforward neural network, and each block reading from and writing to the same residual stream. Figure 1 details only one layer for a single token, although all tokens are processed in parallel. The output of the last layer is mapped back to the vocabulary space by the unembedding matrix $\mathrm{W}_U \in \mathbb{R}^{N\times D}$ and then fed into a softmax layer. Each element in the output vector $\vec{z}_i$ represents the probability of a token being the successor to $\vec{t}_i$.

A zoom into the attention mechanism is provided in Figure 2. It is based on linear algebra operations using the following matrices:

- $\mathrm{W}_K \in \mathbb{R}^{A\times D}$—the "key" matrix;
- $\mathrm{W}_Q \in \mathbb{R}^{A\times D}$—the "query" matrix;
- $\mathrm{W}_V \in \mathbb{R}^{A\times D}$—the "value" matrix;
- $\mathrm{W}_O \in \mathbb{R}^{D\times A}$—the "output" matrix.

A is the dimension of the vector used in the attention computation, in most current NLMs is equal to D. The matrices $\mathrm{W}_{K,Q,V}$ map an embedded token into the vectors "query" $\vec{q}$; "key" $\vec{k}$; and "value" $\vec{v}$. The scalars s_i in Figure 2, called "score", result from the multiplication of the "query" and "key" vectors, and modulate the amount of the "value" vectors.

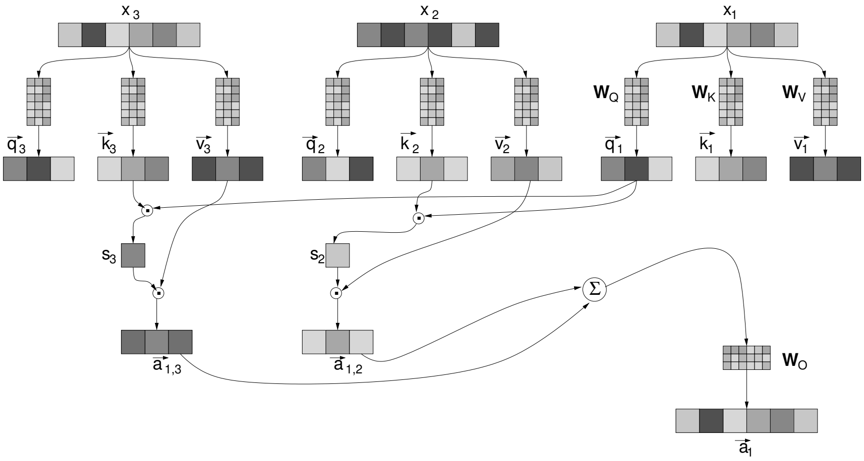

Figure 2. Detail of the attention mechanism, for the current embedded token $\vec{x}_1$ with respect to the previous tokens $\vec{x}_2$ and $\vec{x}_3$.

In a discursive manner, the attention mechanism generates a vector where information from all preceding words is combined, weighted by the relevance of each previous word to the current one. This mechanism synergizes with the other fundamental component of the Transformer: word embedding. The ability to encapsulate all relevant information of a word into a numerical vector for any context of use enables simple linear algebra operations to effectively capture the syntactic and semantic relationships within a text. Now here is the mathematical expression of the operations carried out by the attention:

$$\vec{a}_i = \mathrm{W}_O \mathrm{W}_V \begin{bmatrix} \vec{x}_i \\ \vec{x}_{i+1} \\ \cdots \\ \vec{x}_{i+T} \end{bmatrix} \left(\frac{1}{\sqrt{D}} \begin{bmatrix} \vec{x}_i \\ \vec{x}_{i+1} \\ \cdots \\ \vec{x}_{i+T} \end{bmatrix}^{\top} \mathrm{W}_K^{\top} \mathrm{W}_Q \vec{x}_i \right) \tag{2}$$

where T is the span of tokens preceding the current token $\vec{x}_i$.

The scientific community has been profoundly impacted by the sudden and unforeseen emergence of artificial systems, enabled by Transformer-based models, which exhibit linguistic performances approaching those of humans [20–22,36,37]. For sure, no Transformer-based system matches humans in mastering language in all its possible uses,[2] but the leap made in approaching human performance has been extraordinary. Currently, NLMs continue to progress, whether this means surpassing humans in the near future, or continuing to approach them at an increasingly slower pace [38], is not a matter addressed in this article.

The crucial philosophical issue has become that of providing explanations for the kind of mind that emerges in NLMs and allows its performance, its "alien intelligence" using the words of [39]. Explanations that are currently largely lacking, although some initial attempts can be seen. The almost total absence of explanations for the linguistic abilities of the NLMs contrasts with the relative simplicity of their computational architecture and their way of learning. Again, there is a vast technical literature that computationally illustrates the implementations of the various NLMs [40,41], but there is a huge gap from here to identifying what in these implementations gives language faculty. One of the best illustrative texts on Transformer architectures ([42], p. 71) underscores the issue well: "It has to be emphasized again that there's no ultimate theoretical reason why anything like this should work. And in fact, as we'll discuss, I think we have to view this as a—potentially surprising—scientific discovery: that somehow in a neural net like ChatGPT it's possible to capture the essence of what human brains manage to do in generating language".

Such an explanatory request concerns how the relatively simple algorithmic components of the Transformer provide it with the ability to express itself linguistically and to reason at a level comparable to humans. It's worth noting that while linguistics has generated highly sophisticated and detailed descriptions of language, how it is understood and generated by the brain remains essentially a mystery, much like in NLMs. At the same time, one of the ambitions of simulative AI has been to explain aspects of natural cognition by designing their equivalents. However, the presupposition was that these artificial equivalents would be understandable, which is not the case with NLMs.

Before examining how this challenges the traditional epistemology of simulative AI, let us preliminarily see how NLMs are being used in simulative studies of the brain.

3. Using NLMs to Simulate the Brain

There is a current line of research which investigates the relationships between NLM structures and brain structures, through functional magnetic resonance imaging (fMRI), when engaged in the same linguistic task. It is a surprising inquiry, unexpected even for its own protagonists. Indeed, apart from the generic inspiration from biological neurons for artificial neurons, there is nothing specific in the Transformer mechanisms that has been designed with the brain language processing in mind. However, early results show surprising correlations between activation patterns measured in the models and in the brain, and some analogies in the hierarchical organizations in models and cortex.

Ref. [24] aim at explaining one main difference occurring between NLMs and brain language processing, namely that while NLMs are trained to guess the most probable next word, the brain is able to predict sensibly longer-range words.

Ref. [24], in collaboration with Meta AI, did several experiments to examine correlations between NLMs and brain activities using a collection of fMRI recordings of 304 subjects listening to short stories, and prompting the GPT-2 model with the same stories. Individuals were tested using 27 stories between 7 and 56 min, on average 26 min for each subject, and a total of 4.6 brain recording hours for the 304 subjects. The GPT-2 model involved a pre-trained, 12 layer, Transformer, trained using the Narratives dataset [43].

The first experiment was turned to correlate activations in the Transformer to fMRI brain activation signals for each brain voxel and each individual. Correlations were quantified in terms of a "brain score", determined through a linear ridge regression. In particular, GPT-2 activations linearly mapped on such brain areas as the auditory cortex, the anterior temporal area, and the superior temporal area.[3]

In a second set of experiments, the authors evaluated whether considering longer-range word predictions in the Transformer produces higher brain scores. Longer-range predictions were obtained by concatenating the Transformer activation for the current word with what the authors named a "*forecast window*", that is, a set of w embedded future words, where w is called the width of the window, and where each word is parameterised by a number d, designating the distance of the word in the window with the current word. The experiment yielded higher predictions scores, in this case called "forecast score" (on average +23%) for a range of up to 10 words ($w = 10$), with a peak for a 8 word-range ($d = 8$). Again, forecast score picks correlate model activations with brain activation in cortex areas that are associated with language processing.

In the third, most revealing, experiment, ref. [24] started by the consideration that the cortex is structured into anatomical hierarchies and asked whether different layers in the cortex predict different forecast windows w. In particular, they aimed at evaluating the hypothesis that the prefrontal area is involved in longer-range word predictions than temporal areas. Similarly, the authors considered the different Transformer layers and looked for correlations between activations of the cortex layer and activations of GPT-2 layers. Subsequently, they computed, for each layer and each brain voxel, the highest forecast score, that is, the highest prediction from Transformer layer activations to brain activations. The experiment results were in support of the initial hypothesis.[4]

As stated at the beginning of this section, the work of [24] belongs to a whole line of research looking for correlations between brain structures and NLM structures. To quickly give another example, Kumar and coworkers at the Princeton Neuroscience Institute [26] investigated possible correlations between the individual attention heads[5] in the Transformer, and brain areas when listening to stories. They used the simple model BERT, with 12 layers and 12 attention heads, and applied Principle Component Analysis to the 144 model activations along the story, correlating them with brain areas obtained through fMRI.

What emerges from this line of research, is that Transformer based NLMs are used to model and predict activation patterns in the brain, usually observed through fMRI, in order to collect additional evidence on the brain areas involved in specific linguistic tasks. Schematically, both systems, the NLM and the brain, are given the same task, namely elaborating acoustic signals (the listened story) to process language understanding. The artificial system is then used to predict behaviours (brain activations) of the natural one. This method can be preliminarily considered an instance of the simulative method in AI, that we now turn to analyse.

4. The Simulative Method in Cognitive Science

The *simulative method* in science [45,46] consists in representing a target, natural, system by a means of a mathematical model, usually a set of differential equations, implementing the model in a computational one, typically a simulative program, and executing the latter to provide predictions of the target system behaviours. One characterising feature of computer simulations in science is that they are required to mimic the evolution of the target system in order to provide faithful predictions.

In the realm of cognitive science, the simulative method amounts to implementing an artificial system, either a robot or a computer program, aimed at testing some given hypothesis on a natural cognitive system [47,48]. That is, the main aim of simulations in cognitive science is epistemological: their characterising feature is that they are involved in advancing and testing cognitive hypotheses over the simulated system by building an artificial system and experimenting on it. Experimental strategies are thus performed on the artificial system in place of the natural one. Given a cognitive *function*, hypotheses usually concern the *mechanism* implementing that function in the natural cognitive system.[6] The simulative or, as it is often called, the "*synthetic*" method in cognitive science develops an artificial cognitive system implementing that mechanism for the given function and compares the behaviours of artificial and natural systems. Hypothesised mechanisms play the epistemic role of program *specifications* for artificial computational systems.[7] In case the displayed function of the simulative system matches with the behaviours of the simulated system, the initial hypothesis concerning how the function under interest is realised in terms of the implemented mechanisms is corroborated. Once corroboration is achieved, simulations on the artificial system are used to predict, and explain, the future behaviours of the natural system. Additionally, new mechanisms identified in the artificial system for some displayed function are used as hypotheses for explaining similar behaviours in the natural system.

The synthetic method in cognitive science finds in the *Information Processing Psychology* (IPP) of [52] one important pioneering application. In the approach of Newell and Symon, a human agent is given a problem solving task, typically a logic exercise or the choice of moves in a chess game, asking him to think aloud, thus obtaining a verbal account of her mental processes while carrying out the task. Verbal reports are analyzed in order to identify the solution strategies adopted by the agent and the specific operations performed while carrying out the task. The analysed verbal reports are then used to develop a program that simulates the behaviour of the human agent. Subsequently, new problem solving tasks are given to both the program and the human agent, and verbal reports of the latter are compared with the execution traces of the simulative program to ascertain that the two systems use the same solution strategies. Finally, the program execution traces for new tasks are used for predicting the strategies and mental operations that the human agent performs when given the same tasks.

In the IPP approach, human agents' verbal reports are used to hypothesise the mechanism used by the agents to profitably solve the administered cognitive task. The solution strategies hypothesised by Newell and Symon typically consisted in research mechanisms in decision trees. Research mechanisms of this sort are used as program specifications to develop computer programs, using such programming languages as *Information Processing Language* and *List Processor* (LISP), being able to realise those solution strategies. The *Logic Theorist* and the *General Problem Solver* are well-known examples of such programs. Computer programs are then used to test the initial hypothesis, namely the solution strategy advanced on the basis of the verbal reports. The hypothesis is tested by administering new cognitive task to the program, such as proving logic theorems from Russel and Whitehead's *Principia Mathematica*. In case the solution strategies adopted by the simulative program are the same used by the tested human agent, the initial hypothesis is considered as corroborated.

The synthetic method has been also, and more recently applied, to *biorobotics*. For instance, ref. [53] argue that the synthetic method in simulative AI is the method applied, among others, to the robotic simulation of chemiotaxis in lobsters [54].[8] Ref. [54] hypothesise the biological mechanism implementing lobster chemiotaxis, namely the ability to trace back the source of food, leaving chemical traces in the sea, through chemical receptors put on the two antennae. The very simple advanced mechanism is that the receptor stimulation activates, in a proportional manner, the motor organs of the side opposite to that of the antenna. In other words, the stimulation of receptors of the right antenna activates the left motor organs and the stimulation of receptors of the left antenna activates the right motor organs. The higher the receptor stimulus, the higher the motor organ activation. This simple mechanism would, according to [54], allow lobsters to constantly steer towards the food source following the chemical trail.

Such a hypothesis is tested by building a small robot lobster, named RoboLobster, provided with two chemical receptors, put on the left and right side, and wheels in place of legs. RoboLobster implements the hypothesised mechanism: the left artificial receptor causes, upon stimulation, a directly proportional activation of the right wheel, the right receptor activates the left wheel. RoboLobster was tested in an aquarium containing a pipe releasing a chemical trail. However, the robot was able to trace back the pipe only when put within a 60 cm distance from the pipe; while when put 100 cm away from the chemical source the robot was unable to locate the pipe. The synthetic experiments led the authors to falsify and reject the hypothesis.

Ref. [53] are very careful to notice that when the initial hypothesis gets falsified while testing the artificial system, researchers still use the simulation to understand why the hypothesis was falsified and whether the problem was the hypothesis itself or rather other side phenomena. In other words, they look for an explanation concerning why the supposed mechanism is not able to implement the interested cognitive function. Researchers usually evaluate whether the developed artificial system is a faithful implementation of the hypothesised mechanism. Another source of mistake may be that the mechanism implemented by the developed system is not a faithful description of the biological mechanism.[9]

Ref. [54] suppose that RoboLobster was unable to trace the chemical source because of a wrong distance between the two receptors or of the initial orientation of the robot in the aquarium. However, even modifying the receptor distance and the robot orientation, RoboLobster is still unable to find the pipe when put 100 cm away from it. The authors conclusion is that RoboLobster fails since from a certain distance the chemical trail is scattered and is not informative enough for the robot about the direction to take.

In this third case, the artificial system is used to *discover* new hypothesis about the natural cognitive system and its environment. It is indeed hypothesised that chemical trails are informative with respect to the food location for real lobsters only at a certain distance, the reason being that lobster receptors at a certain distance are not able to detect a difference in chemical concentrations.

To sum up, the synthetic method in cognitive science is a simulative approach applied in all those cases in which testing a cognitive hypothesis directly on the natural system is not feasible. An artificial system is built, in the form of a computer program or robot, and the hypothesis is tested on the artificial system instead. This is done by implementing the hypothesis, in the form of a mechanism for the given cognitive function, in the artificial system and comparing the behaviours of the simulative system with those of the simulated one. In case the artificial system performs the same cognitive function of the natural simulated system, the initial hypothesis is corroborated, otherwise the hypothesis is falsified. In both cases, artificial systems can be used to advance new hypotheses about the behaviours of artificial and natural systems which are tested again on the artificial one. The epistemological relations entertained by the natural cognitive system and the simuative model are depicted in Figure 3.

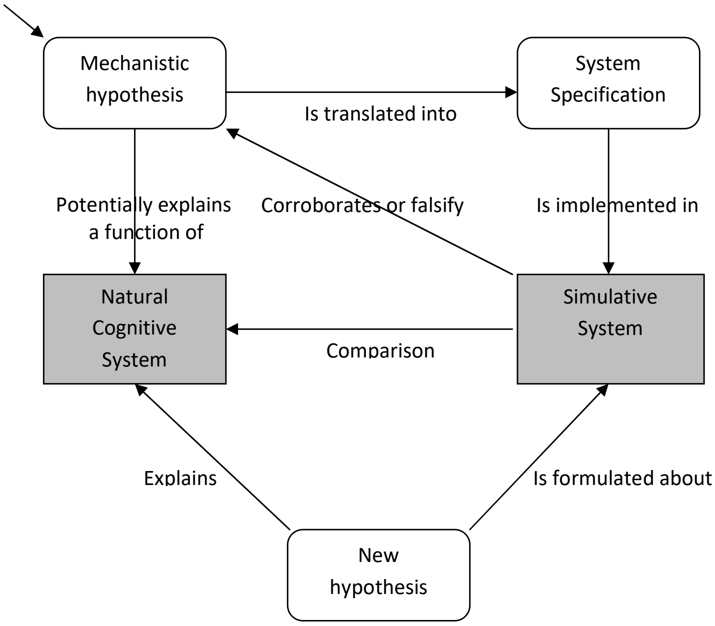

Figure 3. The epistemological framework of simulative AI. The incoming arrow indicates where the process starts.

5. Co-Simulations of Neural Activations Using NLMs

Even though NLMs have been developed with engineering purposes only, namely for developing language processing systems, the early work of [24] and of [26] shows how they are being fruitfully applied to simulative AI as well.[10] However, the way NLMs are used to predict and explain brain activations in the cortex puts significant methodological challenges for the synthetic method in simulative AI.

One first main difference between the simulative method in AI and the application of NLMs in neuroscience is that NLMs are not developed so as to implement mechanisms corresponding to hypotheses about linguistic functions of the brain. The aim of NLMs is not that of corroborating any such hypotheses, as it happens with the simulative method in traditional AI. From an epistemological and methodological point of view, NLMs seem not to be simulative models. And nonetheless, NLMs are used to simulate the brain, that is, to obtain predictions of cortex activations. It is astonishing how, as the work of [24] shows, even though NLMs were developed without considering structural properties of the cortex, once trained they bear structural similarities with language processing areas of brain. An astonishment one also feels while considering DL models involved in vision.[11]

In the synthetic method, hypothesised mechanisms are used as specifications to develop simulative systems and, as stated above, it is required that simulative programs or robots be correct implementations of those mechanisms. As it is in software development, the specification set determines a blueprint of the system to be developed and both correct and incorrect behaviours of the implemented system are defined and evaluated by looking at the specifications [61]. In the case of a correctly implemented system, the specification set provides a means to represent and explain the behaviours of the systems [62]. The opportunity to understand and explain machine behaviours allows scientists to use computational artificial systems for simulating natural ones which, by contrast, are not known and explained.

ANNs in general, and DL models in particular, do not fall under this epistemological framework. DL systems are not developed so as to comply with a set of specifications, that is, functions are not declared and then implemented in a DL network, as it is for traditional software. Functions do not depend only from the network architectural choices, but they rather emerge from the model during training and depend much more on the training dataset [63]. Again NLMs are not developed as implementing neurological mechanisms one supposes realise linguistic functions. The absence of a specification set for NLMs is at the basis of the known *epistemic opacity* of those models: except from some architectural choices (i.e., kind of DL models or the number of models) and hyper-parameters (such as the number of neuron layers or the size of the layers) one is unaware of the inner structure of a trained model. In particular, one cannot come to know how the model parameters are updated at each backpropagation of the network.

In the synthetic method, simulative systems are used as some sort of *proxy* for the simulated cognitive system: since one cannot directly experiment on the cognitive system, as long as it is opaque to the scientist, an artificial system is built and hypotheses are evaluated over it. In the case of Newell and Symon's IPP, since one does not know whether the hypothesised solution strategies for a given task are the ones actually implemented in the brain, the identified research mechanisms for decision trees are implemented in a computer program, the program is subsequently executed to test the hypothesised solution strategies.

The second main epistemological difference of simulations using NLMs is that that they are opaque systems as well and cannot play the epistemic role of proxies for the simulated systems. As what concerns the language function, one is in the difficult situation in which both the natural and the AI system need to be explained. Our knowledge about how the brain processes language is limited in the same way as it is our knowledge about why NLMs show linguistic abilities close to those of humans. As stated in Section 2, such an explanatory gap has been recognized and theorised in one of the most recent technical introduction to NLMs [42].

What the work of [24] shows is that, in front of two opaque systems, they are used to understand each other. As already noted, the simulation starts with no initial hypothesis, being the NLM developed independently from any previous study of brain language processing. Subsequently, and in accordance with the standard synthetic method, both the natural cognitive systems (the 304 tested subjects) and the NLM (GPT-2) are given the same task, namely listening, and processing, 27 short stories, and it is evaluated whether behaviours of the artificial system cope with behaviours of the natural system. In this case, it is tested whether activations in the Transformer can be correlated with fMRI brain activation signals.

Once obtained a positive answer, new experiments are performed to test whether considering longer-range word predictions would decrease the correlation score. One should notice that a hypothesis is involved here, namely that the Transformer differs from the brain while processing language in that the former is able to predict only short-range words, typically the next word in a context. The outcome of the experiment is that the Transformer correlates to the brain more than expected, viz. while predicting up-to-10-range words.

The third experiment is devoted to understand why this is the case, that is, why the initial hypothesis was partly falsified. Notice that this is what happens with the synthetic method too: in case the initial hypothesis gets falsified, further experiments on the simulative system are carried out to understand why this happened. In the case of RoboLobster, once the initial hypothesis concerning the mechanism allowing chemiotaxis was falsified, researches supposed that the inability of the robot to trace back the chemical source, when put on a 100 cm distance, was due to the distance between the two receptors or to the initial orientation of the robot, rather than to the falsity of the hypothesis per se. The robot was tested at different orientations in the aquarium and changing the distance between antennae: experiments were still carried over the artificial system.

Getting back to the GPT-2 experiment, ref. [24] try to evaluate whether the fact that the artificial system and the natural one are both able to predict long-range words can be related to structural similarities between the cortex and the Transformer. This is achieved by considering the cortex *as a model of* the Transformer! In particular, it is hypothesised that the hierarchical organization of the cortex resembles, both structurally and functionally, the hierarchical organization of the Transformer. The hypothesis is tested by administering again the same task to both systems and computing the forecast score, obtaining positive evidence.

When NLMs are used for simulation purposes, one is dealing with a system which is at least as opaque as the natural system about which she would like to acquire knowledge. In the work of [24] the problem is tackled by modifying the simulative approach in such a way that the two opaque systems are used to *simulate each other*, and thus to acquire knowledge about both in the form of corroborated, or falsified, hypotheses. In what can be called a *co-simulation*, the NLM is initially used to simulate the brain by looking for correlations while involved in the same task. In this case, hypotheses to be tested relate to the brain (its ability to predict longer-range words) and correlations are Transformer predictions of brain activations. In case one needs additional information concerning why a certain hypothesis was corroborated or falsified, the natural system is used to simulate the artificial one. Hypotheses now concern the Transformer (its hierarchical organization) and simulations involve brain predictions of Transformer activations. The simulative relations entertained by the brain and NLM are depicted in Figure 4.

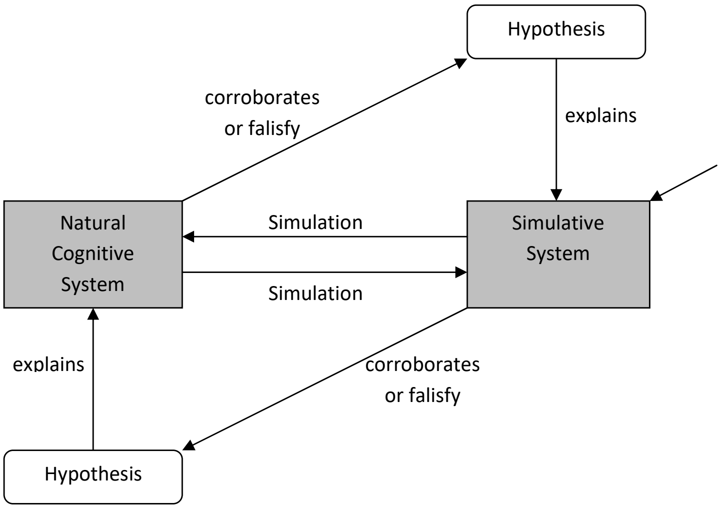

Figure 4. The epistemological framework of NLM simulations.

6. Conclusions

Contemporary DL applications often feature simulation-based scenarios where a model exposed to data from a natural system develops internal structures that correspond to aspects of that system. For instance, ref. [64] utilized a convolutional DL model to simulate parton showers, with each layer representing a different angular scale for emissions. Similarly, in the neural model by [65], which simulates the Hénon-Heiles potential, the autoencoder's internal layer with four neurons captures the four dimensions of the Hénon-Heiles system.

This paper examined another crucial field wherein DL simulations are being applied, namely cognitive neuroscience. NLMs, initially engineered to automatise language translation and generation, are now applied to the simulative investigations of brain language processing. Whereas using artificial computational systems to simulate natural ones is a well-affirmed practice in AI, this paper showed how the applications of NLNs in brain simulations involves significant epistemological and methodological modifications of the synthetic method in cognitive science. The epistemic opacity of NLMs implies that, while they are used to simulate the brain, knowledge is attained about the model as well. This is achieved by a co-simulation wherein the brain is used as a model of the NLM, providing predictions of the Transformer behaviours, and corroborating hypotheses about the latter.

Author Contributions: All authors contributed equally to the conceptualization, methodology, investigation, writing, and reviewing of this work. All authors have read and agreed to the published version of the manuscript.

Funding: Nicola Angius was supported by the PRIN project PRIN_20223E8Y4X_002–CUP J53D23019 490006, "SMARTEST–Simulation of Probabilistic Systems for the Age of the Digital Twin", funded by the Italian Ministry of University and Research. Pietro Perconti, Alessio Plebe, and Alessandro acciai were supported by the PRIN project PRIN_2022MM8KM_003–CUP J53D23007150006, "ALTEREGO–How to Emulate Intentionality and Awareness in Remote Communication by Means of Software Surrogates", funded by the Italian Ministry of University and Research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author/s.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANN	Artificial Neural Networks
DL	Deep Learning
NLM	Neural Language Models

Notes

[1] This will be extensively illustrated in Section 4 below.

[2] A notable case is that no NLM is able to simulate or explain language acquisition by children.

[3] More specifically, the brain score was quantified in the following way. First, a sequence M of words w corresponding to the short stories in the Narratives dataset was defined. The corresponding fMRI recordings from the Narratives Dataset were then sampled with time samples t = 1.5 s and preprocessed using the fMRIprep tool [44] to analyse the cortical voxels; the latter were then projected and morphed onto a brain model, obtaining brain activations Y for each w and having size $T \times V$ (where T is the total number of fMRI samples t and V is the total number of voxels). NLM activations were obtained by tokenising words w in M for being inputted to the network; each activation X corresponded to a vector of size $M \times U$ where U is the number of neurons per layer (768 for the used GPT2 model); activations were mostly extracted from the eighth layer. Finally, for each individual s, each word sequence M, and each voxel v, it was evaluated the mapping between Y and X. The brain score $R^{(s,v)}$ was obtained by using a linear ridge regression to predict a brain activation Y for a given network activation X; the obtained mappings were evaluated using a Pearson correlation between predicted Y and actual activations Y^*. For further technical details the reader should refer to [24].

[4] For technical details the reader should refer to [24].

[5] Embedded vectors in the Transformer are actually divided into portions, called *heads*, and the attention mechanism is applied separately to each head, and only in the end are the various portions re-joined. The idea is that an embedded vector combines different properties of a word, and that certain categories–for example, the tense of verbs or the gender and number of nouns and adjectives–always occupy the same portions of the vector, and therefore it is convenient to process separately the network of relationships between the separate characteristics of the various words in the text.

[6] By mechanism it is referred here to *biological* mechanism as intended in [49], namely as a set of "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination condition" (p. 3). See [50] for how mechanisms of this sort are able to implement cognitive functions.

[7] Program specifications in computer science express the behavioural properties that the system to be developed must realise [51], and their formulation is the first step of most software development methods.

[8] Other biorobotic applications of the synthetic method can be found in the simulation of phonotaxis in crickets [55], ants homing [56], or rats navigation [57].

[9] In the context of the epistemology of computer simulations in science, the two problems are known as the *verification* and *validation* problem for simulative models. Verification is about ascertaining that the simulative system is a correct implementation of the simulative model; validation is about evaluating whether, and the extent to which, the simulative model is a faithful representation of the target simulated system.

[10] It should be indeed recalled that AI has been historically characterised by two main research traditions, an engineering one, concerning the development of artificial systems showing intelligent behaviour, and a simulative one, using artificial intelligent systems to study cognition.

[11] The neuroscience of vision is another field wherein neural architectures keep some feature of the natural system, and important similarities have been found between DL models and the visual cortex [58,59]. DL models have been even found to reproduce structural hallmarks of the visual face network in the inferior temporal cortex [60].

References

1. Wiener, N. *Cybernetics or Control and Communication in the Animal and the Machine*; MIT Press: Cambridge, MA, USA, 1948.
2. Bergson, H. *Creative Evolution*; Dover: New York, NY, USA, 1911.
3. Simon, H.A. *The Sciences of the Artificial*, 3rd ed.; MIT Press: Cambridge, MA, USA, 1996.
4. Tan, K.H.; Lim, B.P. The artificial intelligence renaissance: Deep learning and the road to human-Level machine intelligence. *APSIPA Trans. Signal Inf. Process.* **2018**, *7*, e6. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1090–1098.
6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
7. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
8. Land, E.H.; McCann, J.J. Lightness and retinex theory. *Josa* **1971**, *61*, 1–11. [CrossRef]
9. McCann, J.J. Retinex at 50: Color theory and spatial algorithms, a review. *J. Electron. Imaging* **2017**, *26*, 031204. [CrossRef]
10. Gauthier, I.; Tarr, M.J. Visual Object Recognition: Do We (Finally) Know More Now Than We Did? *Annu. Rev. Vis. Sci.* **2016**, *2*, 16.1–16.20. [CrossRef]
11. VanRullen, R. Perception Science in the Age of Deep Neural Networks. *Front. Psychol.* **2017**, *8*, 142. [CrossRef] [PubMed]
12. Grill-Spector, K.; Weiner, K.S.; Gomez, J.; Stigliani, A.; Natu, V.S. The functional neuroanatomy of face perception: From brain measurements to deep neural networks. *Interface Focus* **2018**, *8*, 20180013. [CrossRef] [PubMed]
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
14. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Kerrville, TX, USA, 2019; pp. 4171–4186.
15. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, S.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
16. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 27730–27744.
17. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Hambro, N.G.E.; Azhar, F.; Rodriguez, A.; et al. LLaMA: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.

18. Alishahi, A.; Chrupała, G.; Linzen, T. Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Nat. Lang. Eng.* **2019**, *25*, 543–557. [CrossRef]
19. Baroni, M. Linguistic generalization and compositionality in modern artificial neural networks. *Philos. Trans. R. Soc. B* **2019**, *375*, 20190307. [CrossRef]
20. Boleda, G. Distributional Semantics and Linguistic Theory. *Annu. Rev. Linguist.* **2020**, *6*, 213–234. [CrossRef]
21. Green, M.; Michel, J.G. What Might Machines Mean? *Minds Mach.* 2022, *forthcoming*. [CrossRef]
22. Pavlick, E. Symbols and grounding in large language models. *Philos. Trans. R. Soc. A* **2023**, *381*, 20220041. [CrossRef] [PubMed]
23. Robinson, L.; Rolls, E.T. Invariant visual object recognition: Biologically plausible approaches. *Biol. Cybern.* **2015**, *109*, 505–535. [CrossRef] [PubMed]
24. Caucheteux, C.; Gramfort, A.; King, J. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nat. Hum. Behav.* **2023**, *7*, 430–441. [CrossRef] [PubMed]
25. Caulfield, J.; Johnson, J.L.; Schamschula, M.P.; Inguva, R. A general model of primitive consciousness. *J. Cogn. Syst. Res.* **2001**, *2*, 263–272. [CrossRef]
26. Kumar, S.; Sumers, T.R.; Yamakoshi, T.; Goldstein, A.; Hasson, U.; Norman, K.A.; Griffiths, T.L.; Hawkins, R.D.; Nastase, S.A. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv* **2023**. [CrossRef]
27. Lipton, Z.C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57. [CrossRef]
28. Rumelhart, D.E.; McClelland, J.L. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; Rumelhart, D.E., McClelland, J.L., Eds.; MIT Press: Cambridge, MA, USA, 1986; Volume 2, pp. 216–271.
29. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
30. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; pp. 3111–3119.
31. Carpenter, G.A.; Grossberg, S. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Appl. Opt.* **1987**, *26*, 4919–4930. [CrossRef] [PubMed]
32. Hinton, G.; Zemel, R.S. Autoencoders, minimum description length and Helmholtz free energy. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 28 November–December 1994; pp. 3–10.
33. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual, 3–10 March 2021; ACM: New York, NY, USA, 2021; pp. 610–623.
34. Milligan, K.; Astington, J.W.; Dack, L.A. Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Dev.* **2007**, *78*, 622–646. [CrossRef] [PubMed]
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
36. Søgaard, A. Understanding models understanding language. *Synthese* **2022**, *200*, 443. [CrossRef]
37. Perconti, P.; Plebe, A. Do Machines Really Understand Meaning? (Again). *J. Artif. Intell. Conscious.* **2023**, *10*, 181–206. [CrossRef]
38. Plebe, A.; Perconti, P. The slowdown hypothesis. In *Singularity Hypotheses: A Scientific and Philosophical Assessment*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 349–365.
39. Frank, M.C. Baby steps in evaluating the capacities of large language models. *Nat. Rev. Psychol.* **2023**, *2*, 451–452. [CrossRef]
40. Tingiris, S. *Exploring GPT-3*; Packt Publishing: Birmingham, UK, 2022.
41. Rothman, D. *Transformers for Natural Language Processing*; Packt Publishing: Birmingham, UK, 2022.
42. Wolfram, S. *What Is ChatGPT Doing ... and Why Does It Work*; Wolfram Media: Champaign, IL, USA, 2023.
43. Nastase, S.A.; Liu, Y.F.; Hillman, H.; Zadbood, A.; Hasenfratz, L.; Keshavarzian, N.; Chen, J.; Honey, C.J.; Yeshurun, Y.; Regev, M.; et al. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Sci. Data* **2021**, *8*, 250. [CrossRef] [PubMed]
44. Esteban, O.; Markiewicz, C.J.; Blair, R.W.; Moodie, C.A.; Isik, A.I.; Erramuzpe, A.; Kent, J.D.; Goncalves, M.; DuPre, E.; Snyder, M.; et al. fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **2019**, *16*, 111–116. [CrossRef]
45. Winsberg, E. *Science in the Age of Computer Simulation*; Chicago University Press: Chicago, IL, USA, 2010.
46. Durán, J.M. *Computer Simulations in Science and Engineering: Concepts-Practices-Perspectives*; Springer Nature: Cham, Switzerland, 2018.
47. Boden, M.A. *Mind as Machine: A History of Cognitive Science*; Oxford University Press: Oxford, UK, 2008.
48. Datteri, E. Biorobotics. In *Agent-Based Modelling in Population Studies: Concepts, Methods, and Applications*; Magnani, L., Bertolotti, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 817–837.
49. Machamer, P.; Darden, L.; Craver, C.F. Thinking about Mechanisms. *Philos. Sci.* **2000**, *67*, 1–84. [CrossRef]

50. Piccinini, G.; Craver, C.F. Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* **2011**, *183*, 283–311. [CrossRef]
51. Turner, R. Specification. *Minds Mach.* **2011**, *21*, 135–152. [CrossRef]
52. Newell, A.; Simon, H.A. *Human Problem Solving*; Englewood Cliffs: Prentice Hall, NJ, USA, 1972.
53. Datteri, E.; Tamburrini, G. Biorobotic experiments for the discovery of biological mechanisms. *Philos. Sci.* **2007**, *74*, 409–430. [CrossRef]
54. Grasso, F.W.; Consi, T.R.; Mountain, D.C.; Atema, J. Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges. *Robot. Auton. Syst.* **2000**, *30*, 115–131. [CrossRef]
55. Webb, B. Robots in invertebrate neuroscience. *Nature* **2002**, *417*, 359–363. [CrossRef] [PubMed]
56. Lambrinos, D.; Möller, R.; Labhart, T.; Pfeifer, R.; Wehner, R. A mobile robot employing insect strategies for navigation. *Robot. Auton. Syst.* **2000**, *30*, 39–64. [CrossRef]
57. Burgess, N.; Donnett, J.G.; Jeffery, K.J.; O-keefe, J. Robotic and neuronal simulation of the hippocampus and rat navigation. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **1997**, *352*, 1535–1543. [CrossRef] [PubMed]
58. Güçlü, U.; van Gerven, M.A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **2015**, *35*, 10005–10014. [CrossRef]
59. Khaligh-Razavi, S.M.; Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **2014**, *10*, e1003915. [CrossRef]
60. Lee, H.; Margalit, E.; Jozwik, K.M.; Cohen, M.A.; Kanwisher, N.; Yamins, D.L.; DiCarlo, J.J. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv* **2020**. [CrossRef]
61. Turner, R. *Computational Artefacts: Towards a Philosophy of Computer Science*; Springer: Berlin/Heidelberg, Germany, 2018.
62. Angius, N.; Tamburrini, G. Explaining engineered computing systems' behaviour: The role of abstraction and idealization. *Philos. Technol.* **2017**, *30*, 239–258. [CrossRef]
63. Angius, N.; Plebe, A. From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning. *Philos. Technol.* **2023**, *36*, 47. [CrossRef]
64. Monk, J.W. Deep learning as a parton shower. *J. High Energy Phys.* **2018**, *2018*, 21. [CrossRef]
65. Choudhary, A.; Lindner, J.F.; Holliday, E.G.; Miller, S.T.; Sinha, S.; Ditto, W.L. Physics-enhanced neural networks learn order and chaos. *Psychon. Bull. Rev.* **2020**, *27*, 217–236. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.