Co-Simulations of Brain Language Processing using Neural Language Models

Nicola Angius (nicola.angius@unime.it)
Department of Cognitive Science, Via Concezione, 6/8
98121 Messina, Italy

Pietro Perconti (pietro.perconti@unime.it)
Department of Cognitive Science, Via Concezione, 6/8
98121 Messina, Italy

Alesio Plebe (alessio.plebe@unime.it)
Department of Cognitive Science, Via Concezione, 6/8
98121 Messina, Italy

Alessandro Acciai (alessandro.acciai@studenti.unime.it)
Department of Cognitive Science, Via Concezione, 6/8
98121 Messina, Italy

Abstract

This paper provides an epistemological and methodological analysis of the practice of using neural language models to simulate brain language processing. 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 examined how the simulative method is modified when using neural language models. In particular, it is argued that the epistemic opacity of neural language models requires that the brain itself be used to simulate the model and to test hypotheses about the model, in what is called here a co-simulation.

Keywords: Philosophy of cognitive science, epistemology of computer simulations, neural language models, language processing, deep learning.

Introduction

Roughly speaking, two main paths can be identified along which the rise of artificial intelligence (AI) has unfolded in the last ten years, driven by the new Artificial Neural Networks (ANN) marked by Deep Learning (DL) (LeCun, Bengio, & Hinton, 2015; Goodfellow, Bengio, & Courville, 2016). 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 (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2015), arousing surprise and interest in the science of vision (Gauthier & Tarr, 2016; VanRullen, 2017; Grill-Spector, Weiner, Gomez, Stigliani, & Natu, 2018).

Five years later, it was the turn of language, a path opened by the Transformer model (Vaswani et al., 2017), quickly followed by various evolutions and variants (Devlin, Chang, Lee, & Toutanova, 2019; Brown et al., 2020; Ouyang et al., 2022; Touvron et al., 2023), 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 (Alishahi, Chrupała, & Linzen, 2019; Baroni, 2019; Boleda, 2020; Green & Michel, 2022; Pavlick, 2023).

The unexpectedly exceptional performance of NLMs has prompted a new line of research in cognitive neuroscience which uses Transformer-based models to simulate brain activities. More specifically, NLMs are used to predict cortex activations while processing language (Caucheteux, Gramfort, & King, 2023; Caulfield, Johnson, Schamschula, & Inguva, 2001; Kumar et al., 2023). Seemingly, this line of research is in continuity with simulative AI, wherein computational systems are developed to simulate human agents involved in some cognitive task. Simulations here involve predictions and explanations of human behaviours.

By contrast, this paper provides an epistemological analysis of NLM applications to the study of brain language processing to argue that significant methodological differences arise with the simulative method as examined in the philosophy of cognitive science. 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, or when the latter is epistemologically opaque. However, NLMs are epistemically opaque as well, since one does not know about the inner structure of the model ones it is trained.

To examine how the simulative method is challenged in NLM simulations, this paper initially introduces NLMs; it shows how they are being used in cognitive neuroscience for simulative purposes; then it recalls what the simulative method in AI is; finally it analyses how the simulative method is applied and modified in NLM simulations.

Neural Language Models

The conquest of natural language has been one of the most difficult challenges for AI, and for a long time, artificial neural networks (ANN) have played a secondary role compared to conventional Natural Language Processing. The initial attempts in this direction (Rumelhart & McClelland, 1986; Elman, 1990) had to confront an apparently irreconcilable gap between the world of language and that of ANN. Language presents itself as a sequence of symbols, directly assimilable

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

in traditional computation, but antithetical to the vectors of real numbers that ANN relies on. The succession of words in the text creates a complex intertwining of semantic and syntactic relationships, the latter well captured by conventional algorithms like parsers, but alien to the static structure of early ANN. A second difficulty that arises from the application of ANNs to natural language processing is that representing words with neural vectors worsens when transitioning from single-word morphology to syntax. Feedforward ANNs are static, and establishing a sense of ordering for multiple words in a sentence are far from straightforward. Moreover, ANN performs at their best when learning in a supervised manner, but in the case of language, there isn't an immediate partition between input and output, on which to conceive a supervised task.

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 (Vaswani et al., 2017) combining several effective strategies to cope with the symbolic nature of natural language. It adopts the *word embedding* technique, which learns from examples the optimal way to transform words into vectors of neural activity (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The numerical vectors can be manipulated with ordinary linear algebra, yielding results that interestingly respect aspects of lexical semantics. The relationship among words in a text is captured by the attention mechanism, initially introduced by Bahdanau, Cho, and Bengio (2016). The Transformer adopts an elegant solution that allows one to bypass supervised learning and which was introduced by (Hinton & Zemel, 1994): the concept of the *autoencoder*. The task assigned to the ANN is simply to reproduce its own output. The architecture that implements it is typically organized into two components. The encoder is responsible for producing an internal representation of the input, and the decoder reproduces the output from this representation, which coincides with the input.

The remarkable efficiency of the Transformer has led to many variations, including ViT *Vision Transformer* (Dosovitskiy et al., 2021) and BERT (*Bidirectional Encoder Representations from Transformers*), where attention is applied to both the left and right side of the current word (Devlin et al., 2019). 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 (Brown et al., 2020). The autoencoding strategy during learning is the task of just predicting the next token in a text. The popular public interface ChatGPT is based on later models of the GPT family (Ouyang et al., 2022).

The sudden and unexpected availability of artificial systems with linguistic performances not so far from human ones offered by Transformer-based models has deeply shaken the scientific community (Boleda, 2020; Green & Michel, 2022; Søgaard, 2022; Perconti & Plebe, 2023; Pavlick, 2023). 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 Frank (2023). 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 (Tingiris, 2022; Rothman, 2022), 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 (Wolfram, 2023, 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 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 AI, let us preliminarily see how NLMs are being used in simulative studies of the brain.

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.

Caucheteux et al. (2023) 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.

Caucheteux et al. (2023), 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 minutes, on average 26 minutes for each subject, and a total of 4.6 brain recording hours for the 304 subjects. The GPT-2 model involved a pretrained, 12 layer, Transformer, trained using the Narratives dataset (Nastase et al., 2021).

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.

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, Caucheteux et al. (2023) 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 in each brain voxel, the highest forecast score, that is, the highest predictions from Transformer layer activations to brain activations. The experiment results were in support of initial hypothesis.[1].

As stated at the beginning of this section, the work of Caucheteux et al. (2023) belongs to a whole line of research looking for correlations between brain structures and NLM structures. To quickly given another example, Kumar and coworkers at the Princeton Neuroscience Institute (Kumar et al., 2023) investigated possible correlations between the individual attention heads[2] 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.

The simulative method in cognitive science

The *simulative method* in science (Winsberg, 2010; Durán, 2018), 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 model, 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 hypotheses on a natural cognitive system (Boden, 2008; Datteri, 2017). 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.[3] The simulative or, as it is often called, the

[1] For technical details the reader should refer to (Caucheteux et al., 2023)

[2] 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.

[3] By mechanism it is referred here to *biological* mechanism as intended in (Machamer, Darden, & Craver, 2000), 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 (Piccinini & Craver, 2011) for how mechanisms of this sort are able to implement cognitive functions.

"*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.[4] 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 Newell and Simon (1972) 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 Theories* 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 programs 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, Datteri and Tamburrini (2007) argue that the syntactic method in simulative AI is the method applied, among others, to the robotic simulation of chemiotaxis in lobsters (Grasso, Consi, Mountain, & Atema, 2000).[5] Grasso et al. (2000) 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 Grasso et al. (2000), 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 author to falsify and reject the hypothesis.

Datteri and Tamburrini (2007) 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.[6]

[4] Program specifications in computer science express the behavioural properties that the system to be developed must realise (Turner, 2011), and their formulation is the first step of most software development methods.

[5] Other biorobotic applications of the synthetic method can be found in the simulation of phonotaxis in crickets (Webb, 2002), ants homing (Lambrinos, Möller, Labhart, Pfeifer, & Wehner, 2000), or rats navigation (Burgess, Donnett, Jeffery, & O-keefe, 1997).

[6] 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.

Grasso et al. (2000) 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 system can be used to advance new hypothesis about the behaviours of artificial and natural systems which are tested again on the artificial one.

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 Caucheteux et al. (2023) and of Kumar et al. (2023) show how they are being fruitfully applied to simulative AI as well.[7] 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 seems 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 Caucheteux et al. (2023) 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 deep neural models involved in vision.[8]

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 (Turner, 2018). In the case of a correctly implemented system, the specification set provides a means to represent and explain the behaviours of the systems (Angius & Tamburrini, 2017). The opportunity to understand and explain machine behaviours allows scientists to use computational systems for simulating natural systems 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 (Angius & Plebe, 2023). 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

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

[8] The neuroscience of vision is another field wherein neural architecture keeps some feature of the natural system, and important similarities have been found between deep learning models and the visual cortex (Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014). Deep learning models have been even found to reproduce structural hallmarks of the visual face network in the inferior temporal cortex (Lee et al., 2020).

program that is then used to test the hypothesised solution strategy.

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 situations 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. Such an explanatory gap has been recognized and theorised in one of the most recent technical introduction to NLMs, namely (Wolfram, 2023).

What the work of Caucheteux et al. (2023) 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 the 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 in involved here, namely that the Transformer differs from the brain while processing language in that the former is ably to predict only short-range words, typically the next word in a context. However, the hypothesis does not concern the simulated system but the simulating one! The outcome of the experiment is that the Transformer correlates to the brain more than expected, *viz.* while predicting to 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, Caucheteux et al. (2023) 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 Caucheteux et al. (2023) 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 relates 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.

Conclusions

Current trends of DL applications involve simulative contexts wherein there is the implicit emergence, inside a model exposed to data of a natural system, of structures that bear some correspondences with structural features of that system. An example is the convolutional DL model used by Monk (2018) to simulate parton shower, where each level of decomposition within the model corresponds to a different angular scale for emissions. And in the neural model by Choudhary et al. (2020), simulating the Hénon-Heiles potential, the internal autoencoder layer with four neurons encodes 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.

Acknowledgments

Nicola Angius was partially supported by the PRIN project PRIN_20223E8Y4X_002 – CUP J53D23019490006, "SMARTEST – Simulation of Probabilistic Systems for the Age of the Digital Twin", funded by the Italian Ministry of University and Research.

References

Alishahi, A., Chrupała, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for NLP: A report on the first BlackboxNLP workshop. *Natural Language Engineering*, *25*, 543–557.

Angius, N., & Plebe, A. (2023). From coding to curing. functions, implementations, and correctness in deep learning. *Philosophy & Technology*, *36*(3), 47.

Angius, N., & Tamburrini, G. (2017). Explaining engineered computing systems' behaviour: the role of abstraction and idealization. *Philosophy & Technology*, *30*, 239–258.

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations.*

Baroni, M. (2019). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical transactions of the Royal Society B*, *375*, 20190307.

Boden, M. A. (2008). *Mind as machine: A history of cognitive science*. Oxford University Press.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, *6*, 213-234.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., & et.al., P. D. (2020). Language models are few-shot learners. *arXiv*, *abs/2005.14165*.

Burgess, N., Donnett, J. G., Jeffery, K. J., & O-keefe, J. (1997). Robotic and neuronal simulation of the hippocampus and rat navigation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1360), 1535–1543.

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

Caulfield, J., Johnson, J. L., Schamschula, M. P., & Inguva, R. (2001). A general model of primitive consciousness. *Journal of Cognitive Systems Research*, *2*, 263–272.

Choudhary, A., Lindner, J. F., Holliday, E. G., Miller, S. T., Sinha, S., & Ditto, W. L. (2020). Physics-enhanced neural networks learn order and chaos. *Psychonomic Bulletin & Review*, *27*, 217–236.

Datteri, E. (2017). Biorobotics. In L. Magnani & T. Bertolotti (Eds.), *Agent-based modelling in population studies: Concepts, methods, and applications* (pp. 817–837). Berlin: Springer-Verlag.

Datteri, E., & Tamburrini, G. (2007). Biorobotic experiments for the discovery of biological mechanisms. *Philosophy of Science*, *74*(3), 409–430.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings north american chapter of the association for computational linguistics: Human language technologies* (pp. 4171–4186). Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations.*

Durán, J. M. (2018). *Computer simulations in science and engineering: Concepts-practices-perspectives*. Cham (Switzerland): Springer Nature.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–221.

Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*, 451–452.

Gauthier, I., & Tarr, M. J. (2016). Visual object recognition: Do we (finally) know more now than we did? *Annual Review of Vision Science*, *2*, 16.1–16.20.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge (MA): MIT Press.

Grasso, F. W., Consi, T. R., Mountain, D. C., & Atema, J. (2000). Biomimetic robot lobster performs chemo-orientation in turbulence using a pair of spatially separated sensors: Progress and challenges. *Robotics and autonomous systems*, *30*(1-2), 115–131.

Green, M., & Michel, J. G. (2022). What might machines mean? *Minds and Machines*, *forthcoming*.

Grill-Spector, K., Weiner, K. S., Gomez, J., Stigliani, A., & Natu, V. S. (2018). The functional neuroanatomy of face perception: from brain measurements to deep neural networks. *Interface Focus*, *8*, 20180013.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Hinton, G., & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In *Advances in neural information processing systems* (pp. 3–10).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1090–1098).

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., … Nastase, S. A. (2023). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv*, *DOI: 10.1101/2022.06.08.495348*.

Lambrinos, D., Möller, R., Labhart, T., Pfeifer, R., & Wehner, R. (2000). A mobile robot employing insect strategies for navigation. *Robotics and Autonomous systems*, *30*(1-2), 39–64.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L., & DiCarlo, J. J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020–07.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–84.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Monk, J. W. (2018). Deep learning as a parton shower. *Journal of High Energy Physics*, *21*.

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., … Hasson, U. (2021). The "narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*, 250.

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

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … et al. (2022). Training language models to follow instructions with human feedback. In *Advances in neural information processing systems* (pp. 27730–27744).

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical transactions of the Royal Society A*, *381*, 20220041.

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

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*, 283-311.

Rothman, D. (2022). *Transformers for natural language processing*. Birmingham (UK): Packt Publishing.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge (MA): MIT Press.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International conference on learning representations.*

Søgaard, A. (2022). Understanding models understanding language. *Synthese*, *200*, 443.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proc. of ieee international conference on computer vision and pattern recognition* (pp. 1–9).

Tingiris, S. (2022). *Exploring GPT-3*. Birmingham (UK): Packt Publishing.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., … Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv*, *abs/2302.13971*.

Turner, R. (2011). Specification. *Minds & Machines*, *21*(2), 135–152.

Turner, R. (2018). *Computational artefacts: Towards a philosophy of computer science*. Berlin: Springer-Verlag.

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, *8*, 142.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 6000–6010).

Webb, B. (2002). Robots in invertebrate neuroscience. *Nature*, *417*(6886), 359–363.

Winsberg, E. (2010). *Science in the age of computer simulation*. Chicago (IL): Chicago University Press.

Wolfram, S. (2023). *What is chatgpt doing ...and why does it work*. Champaign (IL): Wolfram Media.