PNRR - Missione 4, Componente 2, Investimento 1.1 - Bando Prin 2022 - Decreto Direttoriale n. 104 del 02-02-2022
Progetto "Simulation of Probabilistic Systems for the Age of the Digital Twin" (Area CUN 11 S.S.D. M-FIL/02)
CUP J53D23019490006 – codice identificativo PRIN_20223E8Y4X_002

The Simulative Role of Neural Language Models in Brain Language Processing

Nicola Angius [1]
(Work in conjunction with Pietro perconti, Alessio Plebe, Alessandro Acciai)

[1] Department of Cognitive Science, University of Messina, Italy
nicola.angius@unime.it

SMARTEST

SMARTEST 1st Project Meeting
September 10, 2024, CUMO, Noto.

SMARTEST: the UniMe unit

- OBJECTIVE 2: defining the appropriate epistemological framework for simulation in the DT scheme, as it differs from those designed for deterministic computational systems, such as that conceptualised in [Primiero 2019].

    1. TASK 2.1: to verify whether ideal properties of simulation relations hold for the DL scenario;

    2. TASK 2.2: to expand such a framework for DT simulations, wherein industrial artefacts and their digital twins simulate each other in real time.

Angius, N., Perconti, P., Plebe, A., & Acciai, A. (2024). The Simulative Role of Neural Language Models in Brain Language Processing. *Philosophies*, 9(5), 137.

In the Special Issue "Contemporary Natural Philosophy and Philosophies - Part 3 " edited by Gordana Dodig-Crnkovic and Marcin J. Schroeder.

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

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

Why cognitive science?

Computer simulation was pioneered as a scientific tool in meteorology and nuclear physics in the period directly following World War II, and since then has become indispensable in a growing number of disciplines. The list of sciences that make extensive use of computer simulation has grown to include astrophysics, particle physics, materials science, engineering, fluid mechanics, climate science, evolutionary biology, ecology, economics, decision theory, medicine, sociology, epidemiology, and many others. There are even a few disciplines, such as chaos theory and complexity theory, whose very existence has emerged alongside the development of the computational models they study. (Winsberg 2019)

Computer simulations in cognitive science

Cognitive science has been one of the first filed motivating a philosophical reflection on simulation:

- Rosenblueth, A., & Wiener, N. (1945). The role of models in science. *Philosophy of science*, 12(4), 316-321.

- Newell, A., & Simon, H. A. (1961). Computer Simulation of Human Thinking: A theory of problem solving expressed as a computer program permits simulation of thinking processes. *Science*, 134(3495), 2011-2017.

Cybernetics and simulative artificial intelligence

- The use of machines to explain intelligent behaviour traces back to the birth of cybernetics and the work of its founder Norbert Wiener.
- The epistemological setting of cybernetics has been fully inherited by simulative AI, especially in the work of Hallen Newell and Herbert Symon.
- The simulative method of early AI influenced the sub-symbolic paradigm and robotics.
- AI is now living a Renaissance era, thanks to the unexpected success of Deep Learning

The **synthetic method** in cognitive science I

- Advancing and testing cognitive *hypotheses* about a natural cognitive system by building an artificial system and performing a simulation.
- Hypotheses usually concern a *mechanism* implementing a given cognitive function of the simulated cognitive system.
- Simulation amounts to developing an artificial cognitive system implementing that mechanism for the given function and to comparing the behaviours of artificial and natural systems.
- Hypothesised mechanisms play the epistemic role of *program specifications* for the artificial computational system.

The **synthetic method** in cognitive science II

- In case the displayed function of the simulative system matches with the behaviours of the simulated system, the initial hypothesis is corroborated.
- Once corroboration is achieved, simulations are used to predict and explain future behaviours of the natural system.
- New mechanisms identified in the artificial simulative system for displayed functions are used as hypotheses for explaining corresponding behaviours in the natural system.

The Information Processing Psychology of Newell and Simon (1972) I

- A human agent is given a problem solving task, typically a logic exercise or the choice of moves in a chess game, asking her to think aloud, thus obtaining a verbal account of her mental processes while carrying out the task.

- Verbal reports are analysed in order to *hypothesise* the solution strategies adopted by the agent, typically search mechanisms in decision trees.

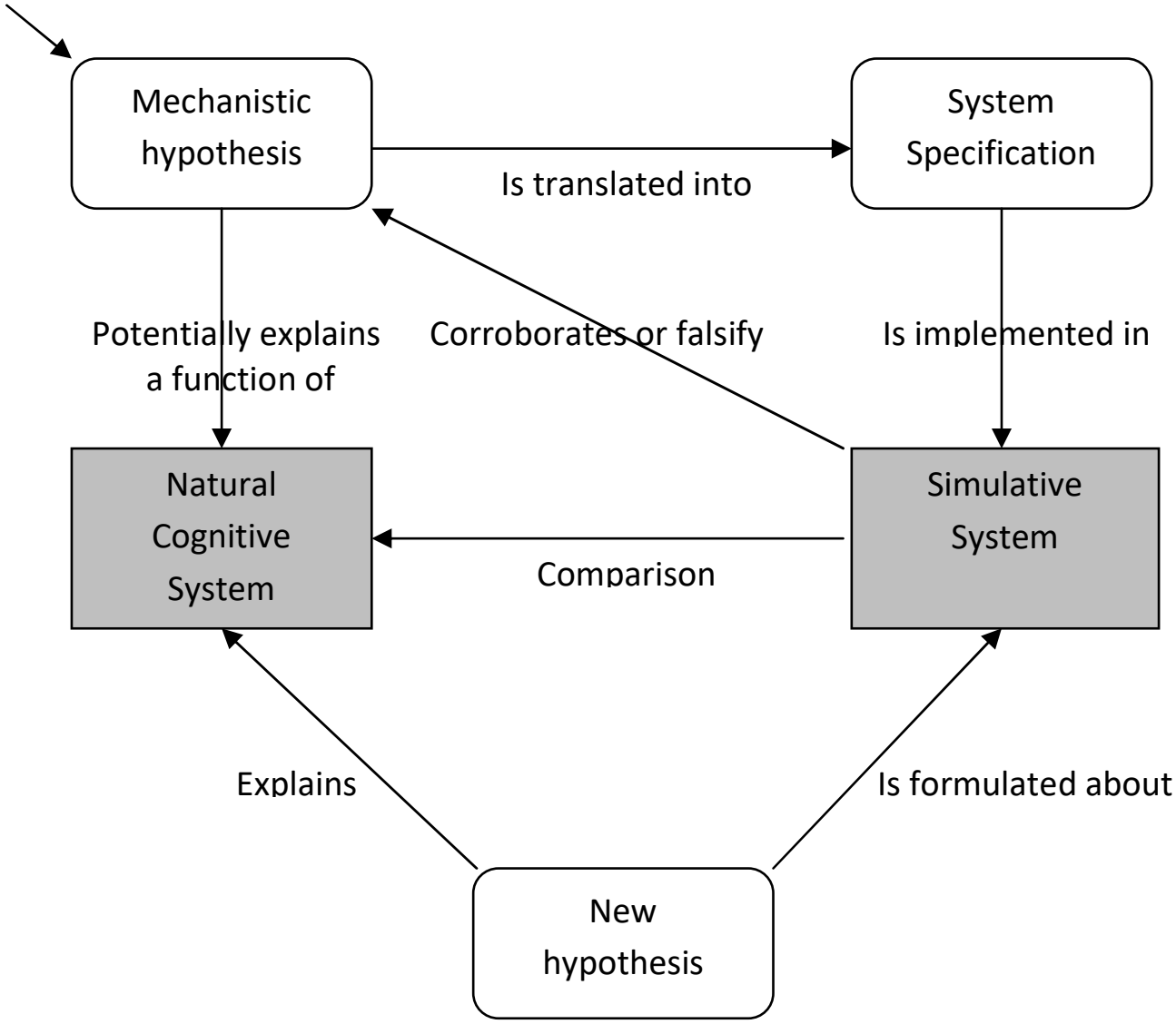
- Search mechanisms are then used as specifications to develop a program that simulates the behaviour of the human agent. (*Logic Theorist, General Problem Solver*)

The Information Processing Psychology of Newell and Simon (1972) II

- New problem solving tasks (proving theorems from *Principia Mathematica*) 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.

The epistemological framework of simulative AI



Deep Learning in Simulative Artificial Intelligence

- Neural Language Models (NLM) are being used for simulative purposes in cognitive neuroscience.
- At the same time, they are *not biologically plausible* and they are *epistemically opaque* (non-interpretable).
- This challenges the traditional epistemological setting of simulative AI, wherein the simulative system is used to understand the mind.

Neural Language Models

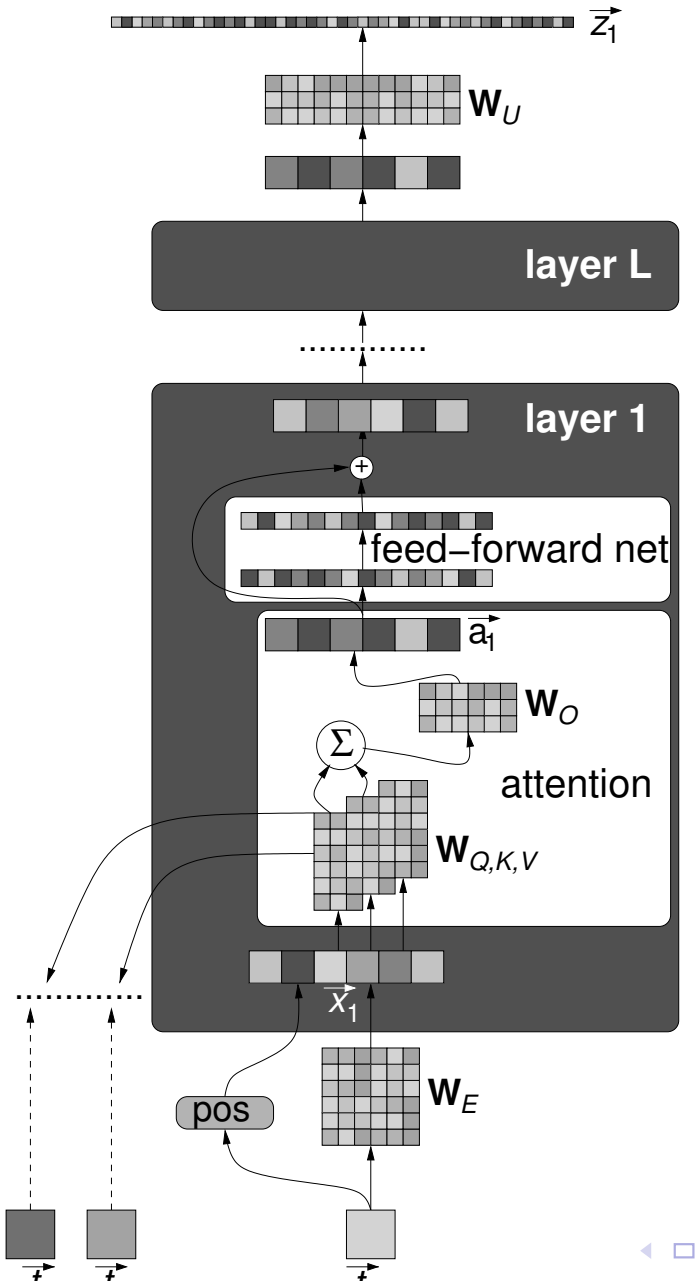
NLMs are deep neural networks based on the Transformer architecture (Vaswani et al. 2017), combining the following strategies to generate language:

1. *Word embedding*, which learns from examples to convert words into, semantically meaningful, real vectors of neural activity (Mikolov et al. 2013).

2. *Attention*, a mechanism identifying relevant information and relationships among words in a sentence (Vaswani et al. 2017).

3. *Autoencoder*, assigning to the network the learning task of reproducing its own input as output (Hinton et al. 1994).

The GPT architecture



Using NLMs to simulate the brain. A study case.

Caucheteux et al. (2023), in collaboration with Meta AI, examined correlations between NLMs and brain activities using a collection of fMRI recordings of 304 subjects listening to short stories, and prompting a GPT-2 model with the same stories.

nature human behaviour

Article

https://doi.org/10.1038/s41562-022-01516-2

Evidence of a predictive coding hierarchy in the human brain listening to speech

Received: 31 March 2022

Accepted: 15 December 2022

Published online: 2 March 2023

Check for updates

Charlotte Caucheteux[1,2], Alexandre Gramfort[1,2] & Jean-Rémi King[1,3]

Considerable progress has recently been made in natural language processing: deep learning algorithms are increasingly able to generate, summarize, translate and classify texts. Yet, these language models still fail to match the language abilities of humans. Predictive coding theory offers a tentative explanation to this discrepancy: while language models are optimized to predict nearby words, the human brain would continuously predict a hierarchy of representations that spans multiple timescales. To test this hypothesis, we analysed the functional magnetic resonance imaging brain signals of 304 participants listening to short stories. First, we confirmed that the activations of modern language models linearly map onto the brain responses to speech. Second, we showed that enhancing these algorithms with predictions that span multiple timescales improves this brain mapping. Finally, we showed that these predictions are organized hierarchically: frontoparietal cortices predict higher-level, longer-range and more contextual representations than temporal cortices. Overall, these results strengthen the role of hierarchical predictive coding in language processing and illustrate how the synergy between neuroscience and artificial intelligence can unravel the computational bases of human cognition.

In less than three years, deep learning has made considerable progress in text generation, translation and completion [1–4] thanks to algorithms trained with a simple objective: predicting words from their nearby context. Remarkably, the activations of these models have been shown to linearly map onto human brain responses to speech and text [5–7]. Additionally, this mapping primarily depends on the algorithms' ability to predict future words [7,8], hence suggesting that this objective suffices to make them converge to brain-like computations.

Yet, a gap persists between humans and these algorithms: in spite of considerable training data, current language models are challenged by long story generation, summarization and coherent dialogue and information retrieval [1–3]; they fail to capture several syntactic constructs and semantics properties [1,9–12] and their linguistic understanding is superficial [10,11,13–15]. For instance, they tend to incorrectly assign the verb to the subject in nested phrases like 'the keys that the man holds ARE here' [16]. Similarly, when text generation is optimized on next-word prediction only, deep language models generate bland, incoherent sequences or get stuck in repetitive loops [11].

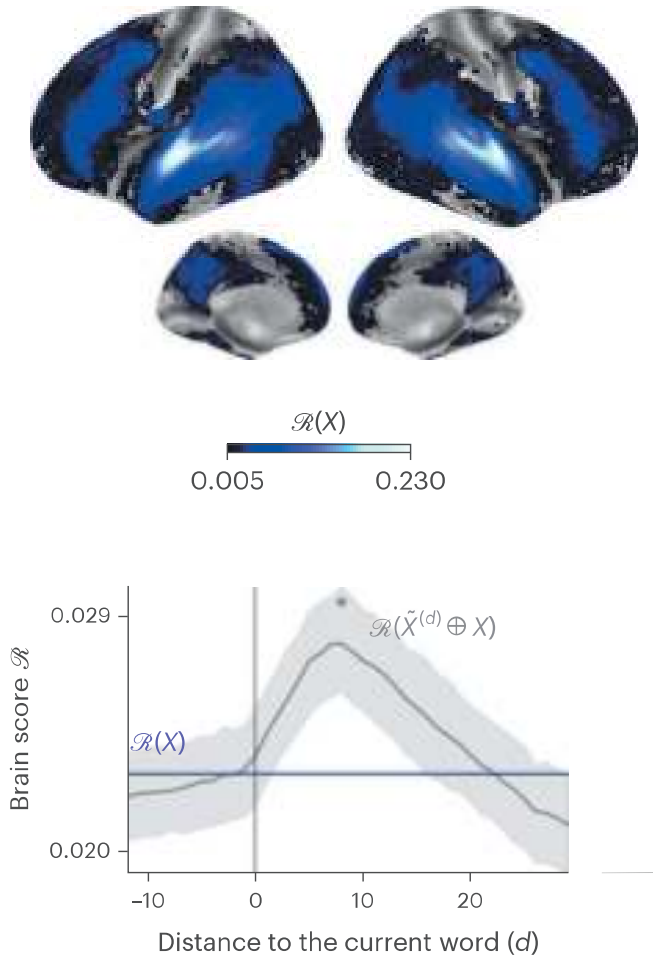
Predictive coding theory [17–21] offers a potential explanation to these shortcomings; while deep language models are mostly tuned to predict the very next word, this framework suggests that the human brain makes predictions over multiple timescales and levels of representations across the cortical hierarchy [22–29] (Fig. 1a).

Previous work already evidenced speech predictions in the brain by correlating word or phonetic surprisal, that is, the extent to which a word or phone is expected, with functional magnetic resonance imaging (fMRI) [30–33], electroencephalography [34–36], magnetoencephalography [37] and electrocorticography [11,38]. However, such surprisal estimates derive from models trained to predict the very next word or phoneme and reduce down their output to a single number, that is, the probability

[1]Meta AI, Paris, France. [2]Université Paris-Saclay, Inria, Commissariat à l'Énergie Atomique et aux Énergies Alternatives, Paris, France. [3]Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, Paris, France. e-mail: ccaucheteux@meta.com; jeanremi@meta.com

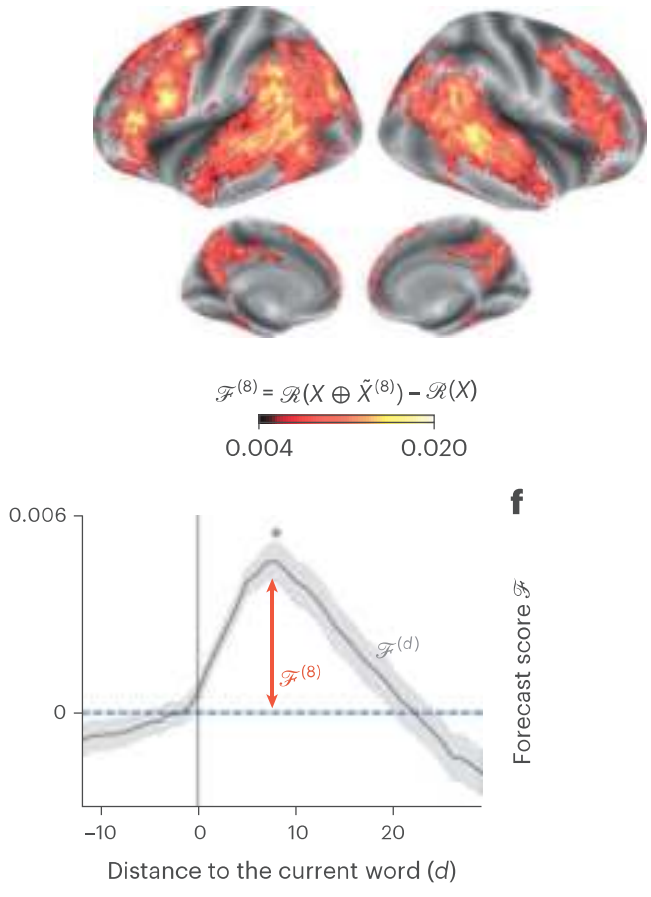
First experiment

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'. GPT-2 activations linearly mapped on such brain areas as the auditory cortex, the anterior temporal area, and the superior temporal area.



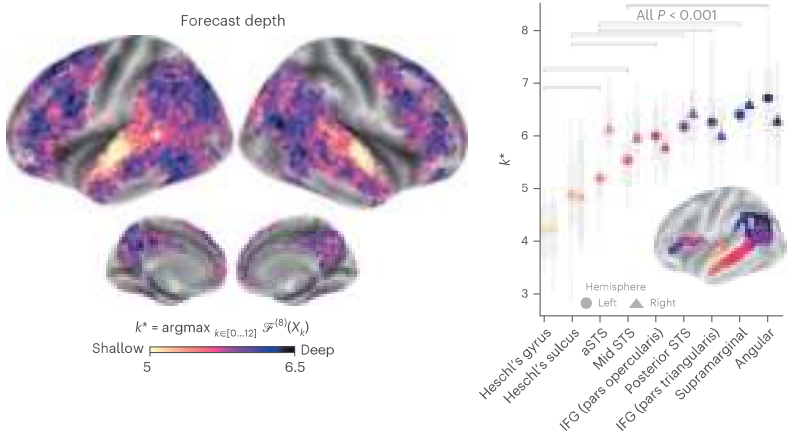
Second experiment

Evaluating whether considering longer- range word predictions in the Transformer produces higher brain scores. The experiment yielded higher (+23%) predictions scores (forecast scores) for a range of up to 10 words, with a peak for 8 word-range.



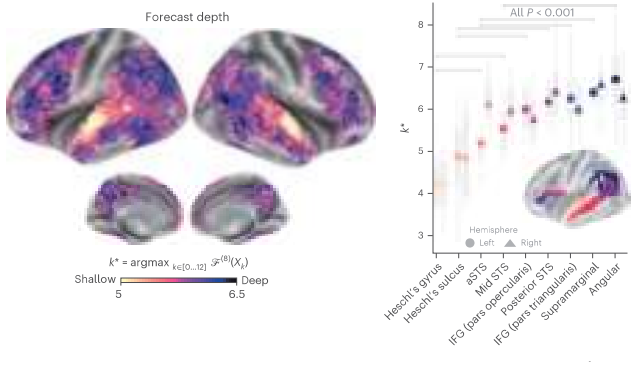
Third experiment I

- Started by the consideration that the cortex is structured into anatomical hierarchies and asked whether different layers in the cortex predict different long-range word predictions (forecast scores).
- Aimed at evaluating the hypothesis that the prefrontal area is involved in longer-range word predictions than temporal areas.
- Considered different Transformer layers and looked for correlations between activations of the cortex layer and activations of GPT-2 layers.



Third experiment II

- They computed for each layer in each brain voxel the highest forecast score.
- The experiment results were in support of initial hypothesis.



NLM simulations

The work of Caucheteux et al. (2023) can be preliminarily considered as an instance of the synthetic method in AI:

- The brain and the NLM are given the same task, i.e. elaborating acoustic signals (the listened story).
- The artificial system is used to predict behaviours (brain activations) of the natural one.

NLM simulations: methodological challenges

The way NLMs are used to predict and explain brain activations puts significant methodological challenges:

1. NLMs are not developed so as to implement mechanisms corresponding to hypotheses about linguistic functions of the brain.

2. DL systems are not developed so as to comply with a set of specifications; functions rather emerge from the model during training and depend much more on the training dataset (Angius and Plebe 2023).

3. NLMs are opaque systems and cannot play the epistemic role of "proxies" of 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.

Instrumentalism or reliabilism?

Most of philosophical reflection on DL simulations is focusing on the epistemic opacity of deep ANNs, swinging between:

- *Instrumentalism*: DL simulations are not to be taken as a scientific method but rather as a tool aimed at enhancing our epistemic capacities (Alvarado 2023).

- *Reliabilism*: DL simulations can be vindicated in case there is some external procedure showing the reliability of the simulation results (Durán and Formanek 2018).

Co-simulations of neural activations using NLMs I

The work of Caucheteux et al. (2023) shows that, in front of two opaque systems, they are used to understand each other in what we call a **co-simulation**:

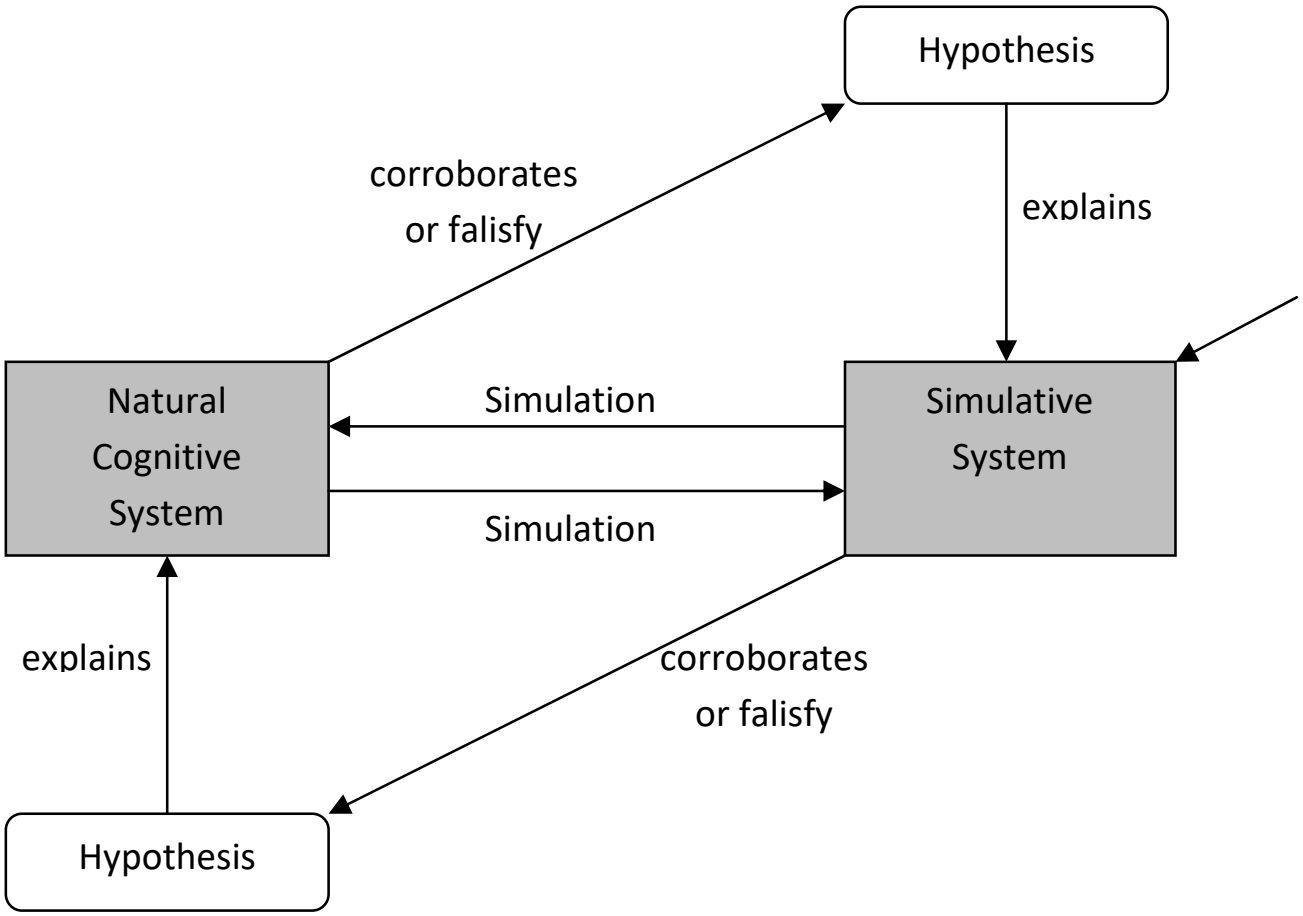
1. In accordance with the standard synthetic method, the natural cognitive systems and GPT-2 are given the same task (listening to stories) and it is evaluated whether behaviours of the artificial system cope with behaviours of the natural system (brain scores).

2. Simulations are performed to **test a hypothesis on the natural system**: the brain is able to predict longer-range words. The hypothesis is falsified by the **Transformer simulation, modelling the brain**.

Co-simulations of neural activations using NLMs II

3. Simulations are performed to **test a hypothesis about the Transformer**: the hierarchical organization of the NLM resembles, both structurally and functionally, the hierarchical organization of the cortex. **The cortex is used as a model of the NLM!** The hypothesis is tested by administering again the same task to both system and computing the forecast score.

The epistemological framework of NLM simulations



Conclusions

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.

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.

A follow-up on task 2.2

- To what extent can the idea of a co-simulation be extended to DT simulations? Is the artefact being developed used to understand its DT?

- Does a DT bear structural, beyond functional, similarities with the industrial artefact? In contrast to NLMs, DTs are developed properly to simulate the artefact.

References

Alvarado, R. (2023). Simulating Science: computer simulations as Scientific instruments (Vol. 479). Springer Nature.

Angius, N., Perconti, P., Plebe, A., & Acciai, A. (2024). The Simulative Role of Neural Language Models in Brain Language Processing. Philosophies, 9(5), 137.

Angius, N.; Plebe, A. From Coding To Curing. Functions, Implementations, and Correctness in Deep Learning. Philosophy & Technology 2023, 36, 47.

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

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds and Machines, 28, 645-666.

Hinton, G.; Zemel, R.S. Autoencoders, minimum description length and Helmholtz free energy. In Proceedings of the Advances in Neural Information Processing Systems, 1994, pp. 3–10

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, 2013, pp. 3111–3119.

Newell, A., & Simon, H. A. (1961). Computer Simulation of Human Thinking: A theory of problem solving expressed as a computer program permits simulation of thinking processes. *Science*, 134(3495), 2011-2017.

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

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, 2017, pp. 6000–6010.

Rosenblueth, A., & Wiener, N. (1945). The role of models in science. *Philosophy of science*, 12(4), 316-321.

Winsberg, Eric, "Computer Simulations in Science", The Stanford Encyclopedia of Philosophy (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2022/entries/simulations-science/>.