Artificial Representations and Intentionality in Machines

Context: If neural networks can be described as possessing representations, what kind are they? How do they compare with the representations of biological systems, including human cognitive systems? This project proposes to explore questions at the crossroads of artificial intelligence, philosophy of mind, philosophy of language and metaphysics. Its aim is to evaluate the ability of certain AIs to represent their environment and to learn from their representational errors.

Artificial representations are most often identified with neural network activation patterns, and are therefore a form of physical information circulating in the computer where the network is implemented. The idea of representation as a network activation pattern has become widespread in AI research. Increasingly, causal intervention techniques are being implemented on artificial neural networks to link the activation of certain neurons to a specific behavior. These techniques involve ablating certain sections of the network to observe changes in behavior and deduce the role of deactivated neurons. Other, more sophisticated methods involve training another AI on the activation patterns of a neural network, in an attempt to predict behavior based on activation. The representations of biological systems, and in particular the mental representations of humans, are commonly described, for their part, as endowed with intentionality (the ability to focus on an object or property) and contents relating to the external world. If mental representation and artificial representation are of the same order, as many authors seem to suggest, this approach to mental representation should be able to be applied correctly to artificial representations. Is this the case?

Scientific objectives: To build well-founded answers to these general questions, four avenues are considered:

(1) Examination of the basis of artificial representations

(2) Examining the possibility of misrepresentation and hallucination in machines.

(3) Semantic capacity and absence of reference: are LLMs capable of producing meaningful statements if their representations do not refer to the outside world?

(1) Founding artificial representations

Mollo and Millière (2023) adapt Harnad's (1990) work on symbol reference in a computational model to connectionist models, and characterize the different ways of grounding a representation. The classic way, the one at the heart of discussions on intentionality, is what they call referential grounding, where the content of a representation is grounded in the object. Referential grounding is a necessary condition for any system endowed with intentional content within a representationalist framework. Indeed, it is necessary for the content of my internal representations to be grounded in the external object in order to consider my mental state as directed towards that object. This is why seeing LLMs as capable of grounding the content of their representations in the external world is a big step in favor of artificial neural networks endowed with intentions. Mollo & Millière give several arguments in favor of referential grounding in certain neural networks, notably those trained by reinformcement learning. I propose to assess the scope of their arguments, and to question their coherence and plausibility. In particular, is it possible to envisage a representation being based on the object, even though the system receives no sensory data? Does the nature of the data influence the representational capabilities of a neural network? Furthermore, relational grounding - a type of grounding in which the content of a representation is based on the relationship between that representation and other representations in the system - seems particularly well suited to connectionist models and LLMs. Should we therefore reject referential grounding in favor of relational grounding?

(2) Possibility of misrepresentation and hallucination

The ability of a system to misrepresent its environment and learn from its mistakes is often considered a major and decisive feature of biological agents. But what about neural networks? Is machine learning a form of misrepresentation correction? If so, against what is the falsity of these representations judged, and is this sufficient to consider that AI learns from the external world in the same way as a biological agent? Does Reinforcement Learning with Human Feedback (Ziegler 2019, Christiano 2017, Ouyang 2022, Liu 2023, Zhang 2023) make it possible to introduce normativity into the training of AIs, by making them confront their representations with certain aspects of the real world? Can we effectively consider, as Hicks (2024) does, that an AI has no relation to truth once it has been trained with human feedback?

(3) Semantic capacity and absence of reference

If artificial representations are not grounded in the external world, how can we interpret the answers given by LLMs? Are they really capable of producing meaningful statements? The distinction between meaning and reference, originally proposed by Frege, remains the subject of much discussion in the philosophy of language. It has been given new impetus by language models. How can we explain the semantic capacities of an LLM if it is incapable of referring to anything and cannot base its representations in the object? For Mollo and Millière (2023), this is a major argument in favor of the intentionality of language models: LLMs' representations are grounded in the external world as part of the chain of reference transmission described by Kripke (1980) and Putnam (1975). Conversely, Bender and Koller (2020) argue that LLMs are nothing more than 'stochastic parrots': since LLMs have no direct causal contact with the world, their representations and syntactic productions cannot refer to anything and are therefore empty of meaning. The publication of Bender and Koller's article sparked off much discussion about the relationship between meaning and reference. Several authors consider that it is not necessary for artificial representations to refer to objects and properties of the world for LLM responses to be interpreted in the same way as human responses (Grindrod 2024, Pavlick 2023, Sogaard 2023). Others opt for an externalist conception in which the meaning of the words produced by a LLM is given by the social community in which the algorithm is embedded (Mandelkern 2024, Lederman 2024).

First, the various proposals made in this debate will be evaluated, and then a conception of the meaning of words based not on reference but on the relations these words have with each other will be examined, based on the article by Piantadosi (2022). This conception seems to provide a satisfactory answer to the question of the meaning of the productions of a language model, without any question of reference or intentionality.
Several related topics might be enlightened by this work.
(a) Agentivity:
Butlin considers that "to be an agent, a system must interact with an environment according to a goal. Intentionality and the ability to represent one's environment may not be strictly necessary conditions for agentivity, but they are at least a sure sign of it.
(b) How the human mind works :
Can the existence of neural networks such as LLMs, to which representations can be attributed, be seen as an argument in favor of a connectionist conception of the mind?
Several authors suggest that the spectacular capabilities of AIs are a clue in favor of connectionism as a global approach to cognition (Buckner 2019).
(c) Using more recent theories of intentional content and theoretical work on learning in the cognitive sciences (Piccinini 2022, Buckner 2021), think about the development of AIs that learn diversely and over longer time scales, so as to extend their causal environment. This broadening of the environment may enable certain models to represent content that is not simply about text or images and does not emerge from the feedback of a small group of socially homogeneous humans (Chaudhari 2024, Christiano 2017), but instead is directly about objects in the world.
Adequation to SCAI: This research project combines elements from philosophy with a strong anchoring in expertiments with LLMs. SCAI apears to be the very best place to develop this research because of its openess to the intellectual risks inherent to interdisciplinary projects.

REFERENCES

Bender, Emily. M., & Koller, Alexandre. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185-5198).
Buckner, C. (2021). A forward-looking theory of content. *Ergo an Open Access Journal of Philosophy*, *8*.
Butlin, Patrick (2024). Reinforcement learning and artificial agency. Mind and Language 39 (1):22-38.
Chaudhari, Shreya et al. (2024). RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *arXiv preprint arXiv:2404.08555*.
Christiano, Paul F., et al. Deep reinforcement learning from human preferences. (2017) *Advances in neural information processing systems* 30.
Frege, Gottlob (1892). Uber Sinn und Bedeutung. Zeitschrift für Philosophie Und Philosophische Kritik 100 (1):25-50.
Grindrod, Jumbly (2024). Large language models and linguistic intentionality. Synthese 204 (2):1-24.
Harnad, Stevan (1990). The symbol grounding problem. Physica D 42:335-346.
Hicks, Michael Townsen ; Humphries, James & Slater, Joe (2024). ChatGPT is bullshit. Ethics and Information Technology 26 (2):1-10.
Kripke, Saul A. (1980). Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium. Cambridge, MA: Harvard University Press. Edited by Darragh Byrne & Max Kölbel.
Lederman, Harvey & Mahowald, Kyle (2024). Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs. Transactions of the Association for Computational Linguistics 12:1087-1103.
Liu, Hao, Carmelo Sferrazza, and Pieter Abbeel. (2023) Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*.
Mandelkern, Matthew, & Linzen, Tal. (2024). Do language models refer?. *Computational Linguistics*, *50*(3).
Mollo, Dimitri Coelho, & Millière, Raphaël. (2023). The vector grounding problem. *arXiv e-prints*, arXiv-2304.
Ouyang, Long, et al. Training language models to follow instructions with human feedback. (2022) *Advances in neural information processing systems:* 35 27730-27744.
Pavlick, Ellie. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, *381*(2251), 20220041.
Piantadosi, Steven T., & Hill, Felix. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.
Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics*, *16*, 846979.
Putnam, Hilary (1975). The meaning of 'meaning'. Minnesota Studies in the Philosophy of Science 7:131-193.
Schulman, John et al. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
Søgaard, Anders (2021). Grounding the Vector Space of an Octopus: Word Meaning from Raw Text. Minds and Machines 33 (1):33-54.
Zhang, Tianjun, et al. (2023) The wisdom of hindsight makes language models better instruction followers. *International Conference on Machine Learning*. PMLR.
Ziegler, Daniel et al. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.