Representational Explanation for Deep Artificial Neural Networks: prospects and challenges

The increase in size and capabilities of current deep artificial neural networks (DNNs), and especially of Large Language Models (LLMs), has turned them more and more into explanatory targets in their own right, instead of being seen as 'just' useful tools for assisting in practical and scientific goals. Consequently, the project of AI interpretability, i.e. the scientific field dedicated to explaining and understanding the functioning of AI systems, has grown considerably in the last 5 years or so.

AI interpretability employs a variety of methods to investigate the behaviour and inner workings of AI systems, many of which are inspired or directly copied from methods in cognitive science, such as behavioural testing, ablations, study of stimulus response profiles, and functional modelling. This methodological adoption brings along with it a baggage of conceptual assumptions and explanatory constraints to AI interpretability that have their origin and justification in the practices and aims of cognitive science.

A key, foundational element in mainstream cognitive-scientific methods and explanations is the appeal to internal representations. Such an appeal also permeates AI interpretability work, although its explanatory fruitfulness is debated. In this talk, I examine the prospects for theories of representation for AI systems. I remain neutral on whether such systems should count as cognitive.

Philosophers of cognitive science have dedicated considerable effort to showing that internal representations can be *bona fide* scientific posits by developing naturalistic theories of representation for biological cognitive systems. Even though there is no consensus on the details of the correct theory of representation, there is wide agreement on its main components. Internal representations are physical states that stand in causal-informational relations to states of the world, and that have the function to carry information about the world, having been selected by natural selection processes to do so.

I identify three main options for theories of internal representation in current AI systems:

a) our best theories of internal representation fully apply to current AI systems;
b) our best theories of internal representation apply to current AI systems, but require important tweaks;
c) theories of internal representation for current AI systems are substantially different from theories of internal representation for biological systems.

I argue that a) is unlikely to be a promising route, that b) is plausible but the nature, extent, and tenability of the required tweaks remain unclear, and that c) is a potentially philosophically fruitful route, but that it may end up collapsing into b) or leading to a rejection of the use of cognitive science methods in AI interpretability research.

In brief, two main challenges plague strategy a). First, the AI systems to which we apply cognitive-scientific methods today are almost exclusively disembodied pieces of software that lack direct causal-informational relations with the world. The data on which they are trained are, instead, human-produced representations, such as texts and photos. Second, it is unclear that current AI systems undergo the selection processes our best theories of representation require. They do not undergo natural selection, the process that most theories give pride of place to. It is moreover unclear whether DNN training count as genuine learning, or at least the kind of learning that can ground representational functions.

When it comes to b), it seems that at least two key tweaks are required. A theory of internal representation for current DNNs should not require direct causal-informational relations to the world, but instead show that exclusively indirect, human-mediated causal-informational relations are sufficient. This is far from a trivial endeavour, as it needs to be shown that AI representations can be about the world despite this mediation, rather than being 'just' representations of human representations. In addition, the theory would need to appeal to non-natural selection processes, perhaps learning-like processes during training. However, an analogous difficulty appears here: even conceding that appropriate selection processes are at play, it must also be shown that the selection 'rationales' are worldly states, rather than purely human representations of such states. Should strategy b) fail to meet these twin challenges, representational explanation for AI systems would at best be fundamentally different to cognitive science explanations: AI representations would be exclusively representations of certain human populations' representational space, rather than representations of the world – with consequent differences in our understanding of AI systems and of whether they can count as cognitive.

Finally, strategy c) is partly motivated by the difficulties pointed out above: the differences between biological systems and DNNs are such that we may need fundamentally different theories of representation for AI systems. On this view, DNNs are a 'new kind of beast', sharing some core features with biological cognitive systems, while lacking others. It is unclear what factors could constitute the grounds for such alternative, DNN-focused theories of representation. If they also end up appealing to causal-informational relations and selection processes, the strategy risks collapsing into b). If, on the other hand, such theories are indeed fundamentally different to theories of representation for biological systems, they run the risk of falling outside the purview of the explanatory practices and methods of cognitive science, thus putting pressure on the extension of cognitive-scientific methods to the study of AI systems. That said, the proof of the pudding is in the eating. Philosophers of AI should try and develop candidate DNN-focused theories of representation, so that we can assess their tenability and independent scientific fruitfulness.

In this talk my job has been mostly that of presenting a space of possibilities and their associated challenges, but it is also a plea for trying out new, unbeaten paths in our philosophical exploration of representation and cognition over and beyond biological systems.