Doing without etiological functions in AI metasemantics

What, if anything, do the internal states of AI models represent? Do activation patterns in large language models, for instance, represent familiar entities like grapefruits and governments, or merely linguistic objects like words and syntactic structures? Recently, philosophers have turned their attention to these questions, attempting to articulate the conditions under which AI internal states bear particular representational contents. An influential approach (taking cues from teleosemantics) makes appeal to *etiological functions* – functions deriving from the causal history of a system or lineage of systems (Butlin 2023; Coelho Mollo & Millierè forthcoming; Goldstein & Levinstein 2024). In this paper, I argue that an alternative kind of function – one deriving in part from the intentions of designers, deployers and users – is better suited for the task of grounding representational content in AI systems.

The etiological approach to AI representation draws on theories originally developed for biological cognition (Milikan 1984; Shea 2018). On standard teleosemantic theories, the representational content of an internal state is the worldly condition that the state has the function of carrying information about. These theories standardly employ an etiological notion of function: the function of an internal state derives from the properties it was *selected for* (e.g. by natural selection), or that were the target of some historical *stabilisation process* (such as feedback-based learning) (Garson 2016; Shea 2018). For instance, if activity *F* in the brains of ancestral frogs contributed to frog proliferation by carrying information about flies, then activity F in present-day frogs has the *function* of carrying information about flies, and in turn, can be said to *represent* flies.

Applied to AI, advocates of the etiological approach have argued that machine learning processes (such as supervised learning or reinforcement learning) can constitute the relevant function-conferring selection or stabilisation processes. On this view, internal states that were selected during training for carrying information about model-external features thereby acquire the function of representing those features.

While etiological accounts of function are dominant in theorising about biological functions and biological representation, the invocation of etiological functions in AI metasemantics has not been sufficiently justified. Do the functions that determine representational content in AI systems stem from the etiology of those systems (e.g. their training history) or are they the result of some other function-conferring factor? One reason to question the assumption of the etiological approach is that AI systems are not biological organisms but *artifacts*, produced, deployed and used by intentional agents for our purposes. So, perhaps human intentions play some role in fixing the (representational) functions of their components.

I identify three desiderata for an account of functions in the context of grounding representational content in AI systems. I then develop an account of intention-derived functions in AI systems, which I dub "deployment functions" and show that they better satisfy these desiderata, compared with etiological functions.

First, I define *system-level deployment functions*: an AI system has a deployment function to *X* if agents design, deploy, or use it to *X*. A chatbot deployed for answering factual questions thus has the (system-level) deployment function of answering factual questions, regardless of whether it was trained on some other task, such as next-token prediction.

Second, I show how system-level deployment functions confer functions on internal components. An initial challenge to developing an intention-based account of functions for components in AI systems is that, unlike ordinary artifacts, the internal workings of deep learning systems emerge

through training and are not hand-coded by engineers. Hurshman (2024) thus argues that components of opaque neural networks lack intention-derived functions, since engineers lack the necessary beliefs about component-level mechanisms. However, I argue that intentions can ground component functions *indirectly*, even when designers are ignorant of component-level mechanisms.

On my account, a component has a *component-level deployment function*, *F*, if it contributes to the system's (system level) deployment function *G* by *F*-ing. Crucially, this does not require that designers have any beliefs about the component. Consider ancient Roman concrete, which is unusually durable in seawater. It turns out that – unbeknownst to the ancient Romans – this is due to specific mineral constituents which form fracture-resistant plate-like structures (Jackson et al. 2017). The Romans could not conceive of these micro-structural properties. Yet the minerals, I contend, served the *function* of forming such structures, because this property is what contributes to the concrete's intended purpose (resistance to seawater exposure). Similarly, components in opaque neural networks can have deployment functions fixed by their contribution to system-level deployment functions, even when engineers are unaware of what those components do, and independently of any historical facts about the training of the system.

This account of deployment functions meets key desiderata for a theory of representational functions for AI systems:

(i) Unlike the related, but weaker notion of a "Cummins function" or "role function" (Cummins 1975; Craver 2001), it exhibits *normativity*: because the system-level function is genuinely normative (the system is *supposed to* behave a certain way), components that fail to make their characteristic contribution to that function are malfunctioning, not merely playing a different causal role.

(ii) It accommodates *misrepresentation*: a component that has the deployment function of carrying information about *P*, but fails to do so, thereby fails to contribute to the system's deployment function and so misrepresents.

While etiological functions also fulfil the above two desiderata, deployment functions satisfy a third desideratum that etiological functions do not: (iii) *explanatory relevance*. Interpretability researchers attribute representational contents to AI model-internals in the service of broader pragmatic goals, such as predicting failure modes, mitigating dysfunctional behaviour, and building models that better serve our interests (Sharkey et al. 2025). These concerns are largely *ahistorical* and impose *use-centred success criteria*: by positing internal representations, researchers typically want to explain how components contribute to the model's performance on the tasks we currently use them for, rather than the (possibly divergent) roles stabilised through training. Deployment functions are thus more faithful to the explanatory goals and practices of AI interpretability research, and should be preferred over etiological accounts in AI metasemantics.

References:

Butlin, P. (2023). Sharing Our Concepts with Machines. *Erkenntnis*, *88*, 3079–3095. https://doi.org/10.1007/s10670-021-00491-w

Coelho Mollo, D., & Millière, R. (forthcoming). The Vector Grounding Problem. *Philosophy and the Mind Sciences*.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, *72*, 741–765.

Craver, C. F. (2001). Role Functions, Mechanisms, and Hierarchy. *Philosophy of Science*, *68*(1), 53–74.

Garson, J. (2016). *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.

Goldstein, S., & Levinstein, B. A. (2024). Does ChatGPT Have a Mind? arXiv preprint: https://arxiv.org/abs/2407.11015v1

Hurshman, C. (2024). Do opaque algorithms have functions? *Synthese*, *204*(3), 1–26. https://doi.org/10.1007/s11229-024-04745-2

Jackson, M. D., Mulcahy, S. R., Chen, H., Li, Y., Li, Q., Cappelletti, P., & Wenk, H. R. (2017). Phillipsite and Al-tobermorite mineral cements produced through low-temperature water-rock reactions in Roman marine concrete. *American mineralogist*, *102*(7), 1435-1450.

Millikan, R. G. (1984) *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., … McGrath, T. (2025). *Open Problems in Mechanistic Interpretability*. 1–82. arXiv preprint: http://arxiv.org/abs/2501.16496

Shea, N. (2018). *Representation in Cognitive Science*. Oxford: Oxford University Press.