Extended abstract for "Can Swampmodels have Inner Representations?" (aka Synchronic Functions for Representation)

Representation helps us to explain complex behaviors, and to make sense of our internal states. Our explanatory goals also often push us towards attributing representations to non-human systems (animals or machines) that exhibit sophisticated capacities.

But what is representation, and how does it fit into a scientific picture of the natural world? How do we find it, measure it, and so on? One strand of philosophical work in this area seeks to *naturalize* representation – by reducing it to other things that seem better understood. Millikan (1984), Dretske (1988) and others sought to explicate representation in terms of processes such as learning and evolution. Unfortunately, these processes are *historical*, and so their accounts of the nature of representation depend on how a system came to be, not just how it works right now. This sits uneasily with many people's philosophical intuitions about cases such as Swampman, as well as with scientific practice in cognitive science, psychology, and neuroscience, which posit representations freely with no explicit reference to evolutionary history. It seems that the concept of representation that we actually want is *synchronic* – that is, ahistorical – but our naturalistic account of it is not. Forty years later, this still seems like a problem (see Grush (2004) for a compelling statement of the problem for computational neuroscience in particular).

I attempt to provide a synchronic notion of function as one of the key ingredients for this project of naturalizing representation. The basic idea is that what matters is not the history of selection itself, but rather the forces or mechanisms that are responsible for that history. Evolution by natural selection under certain conditions is a *stabilizing* process – and so too is learning. It pushes the system towards being one way, and away from other ways of being. At least under certain conditions, once an organism is at a peak in its fitness landscape, it tends to stay there. (Many caveats apply, and much work has been done on exceptions such as frequency-dependent selection, or neutral selection and so on, which may even be more common than the rule. But no one denies that a good approximation to this cartoonish version of adaptation happens often). So what matters is not the particular trajectory that an organism may have taken to its current form, but rather that there were forces shaping that trajectory in a particular way. Making those forces explicit is another way of fleshing out the contribution of evolution and learning to making our representations what they are.

In the teleosemantic tradition, functions are what traits have, in fact, historically been selected to do. According to the closely related synchronic proposal, functions are what traits are *under selection pressure* to do. Selection pressure is determined by the fitness landscape, and that landscape at any one moment is determined by facts about the environment, just as an electromagnetic field is determined by electromagnetic sources in classical EM. Similar ideas about robustness, dynamical convergence and the close relationship between the language of *purpose* and stability have been put forth by Nanay (2010), Birch (2012), Trestman (2012), and Shea (2018).

This overall structure generalizes to learning, as Dretske (1988), Garson (2017), and others have proposed. A learning organism is one that shapes itself so as to be more likely to achieve its goal. A stabilizing process and associated loss landscape is also immediately analogous to the way that machine learning formalizes its problem of training artificial neural networks to produce interesting and sophisticated behaviors. There, an objective function takes the place of the fitness function, and some version of gradient descent takes the place of the mechanism of differential reproduction or retention

that drives evolutionary change. In all cases, we can use the concept of a generalized energy landscape, with some kind of hill-climbing or optimization process pushing systems towards local maxima. The behavior and traits of systems at these local maxima give us the functions of traits of systems in the nearby landscape more generally.

Taking these synchronically defined functions, we can go back to the project of naturalizing representation with an alternative for those unwilling to bite the Swampman bullet, and one closer perhaps to scientific practice as well.