**Sketch for a non-reductive theory of natural agency**

The free will problem has traditionally been understood as one that is specific to human action. One possible reason for this is the widespread interpretation of free will as the control condition on moral responsibility, and where the attribution of moral responsibility is treated a distinctively human phenomenon. However, Steward (2012) has recently argued that this problem is not merely one for human action but for agency wherever we find it in the natural world. That means that there is not only a tension between determinism and human action (or a particular class of human actions), but also a tension between determinism and action *per se*.

This problem derives from two independently plausible views. The first is a metaphysical view of nature that is often associated with (but not necessitated by) contemporary natural sciences. On this view, causation is a relation that holds exclusively between events. Paradigmatic in this regard is a flying ball breaking a window. On the event-causal view, it is not correct to say that the ball broke the window, but rather that the ball's collision with the window caused the window to break. The second is a view of action as comprised as an agent's directly bringing about some state of affairs by means of a basic action (e.g., a bodily movement). These are in tension because an agent is not an event but rather a kind of entity, and entities do not directly bear causal powers on the metaphysical view just mentioned.

Philosophers of action since Davidson have largely dealt with this problem through constructing reductive theories of agency that make sense of action in terms of event causation alone. Davidson (1963) himself, for example, gives a theory on which a behavioural happening counts as an action in cases where it is caused by the agent's having a belief-desire pair that makes the happening rationally intelligible. Such an event-causal theory of action has become the orthodoxy in philosophy of action, despite continuing disagreement over the details (for instance, over what kind of psychological events are the right causal antecedents for making a behaviour an intentional action).

There are reasons why we should prefer a non-reductive theory of agency if one is available. First, it is not clear that the reductive theory preserves an essential feature of action: that it is performed by an agent. Many authors have pointed out that understanding action and its antecedents in terms of events occurring within an agent seem to make the agent vanish altogether. But the agent herself is essential to our quotidian practices of action-explanation and responsibility ascription. We do not punish mental states for bad deeds; we punish agents. Secondly, reductive theories of action are prone to the problem of causally deviant chains (where the behaviour is caused by the relevant mental states but for which the behaviour seems unintentional), which were first flagged by Davidson himself and which remain in want of a satisfactory solution. However, theorists persist with reductive theories due to the presumed philosophical impropriety of agent-causation.

The aim of this paper is to establish the plausibility of a naturalised account of agent-causation, of just the kind required for a non-reductive theory of agency. The account is constructed on the basis of a theory of biological self-organisation called the free energy principle (FEP; see Friston 2013 and references therein). There is a rapidly growing literature in cognitive neuroscience that adopts FEP (often under the name "predictive processing") in order to explain a wide range of perceptual, cognitive, and behavioural phenomena. Despite the fact that it is more commonly discussed in cognitive neuroscience, FEP is nonetheless often motivated by appeal to biophysical considerations regarding very simple biological systems and the physical constraints on these far-from-equilibrium systems when they resist dispersion into the environment. In the rest of the abstract I give a brief primer on FEP and explain why it suggests a way in which to develop a naturalised theory of agent-causation.

Biological systems are called "far-from-equilibrium" in order to highlight the fact that physical systems usually disperse into their environments until a point of equilibrium is reached. Consider a lump of iron that is left in a damp environment. Left alone, the iron will rust away until it has entirely dispersed into its environment and is no longer recognisable as a distinct system. In these cases, what we observe is the interaction of the system and its surroundings in accordance with physical laws until they reach an equilibrium. Most importantly, the point of equilibrium will be one in which the entropy is as low as possible for both systems. That is why we often see dispersion in these cases: the interactions between the system and its surrounds involve an increase in disorder which degrades the distinction between the two. Biological systems appear to be different, because they resist this widespread tendency towards disorder for as long as they survive. This has long been noted as an interesting feature of living things (Schrödinger 1944), and it is why they are called far-from-equilibrium systems: they are much more highly ordered than their immediate surrounds and resist the gradient to thermodynamic equilibrium.

The FEP assumes that i) the behaviour of biological systems is ergodic, and ii) that there are a set of states that partition the organism from its environment, and which are statistically dependent either on the organism (active states) or the environment (sensory states). On the basis of these two assumptions, Friston (2013) has shown that as long as it survives, the organism embodies an implicit model of the environmental factors which cause changes in its sensory states. This is to be construed as a hierarchical generative model that minimises a quantity equivalent to information-theoretic surprise, both by updating its parameters (a kind of proto-sensation) and by sampling the world through changes in its active states to avoid surprising sensory states (a kind of proto-action).

This theory of self-organisation is germane to a non-reductive theory of natural agency in virtue of three features: i) it reveals action to be a phenomenon which is common to all biological systems, thus part of the natural world; ii) it primitivises action, firstly in the sense that it does not reduce action to a species of event-causation and secondly in the sense that it breaks the link between action and

sophisticated cognition; and iii) despite the fact that agency remains unreduced to a species of event-causation (it is rather a relation between an embodied model and some of its states), it nonetheless explains the emergence of agency on the basis of physical principles, which means that it fits the bill for a naturalised metaphysical theory according to the criteria discussed by Ladyman and Ross (2007). (That is not necessarily to endorse their account, but rather to choose a particularly demanding set of standards.) My paper will elaborate on these virtues at length, and defend the account I sketch against potential objections.

**References:**


Davidson, D. (1963). Actions, reasons, and causes. *Journal of Philosophy*, 60, 685-700.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86).

Ladyman, J., & Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

Schrödinger, E. (1944). *What is Life? The Physical Aspect of the Living Cell*. Cambridge: Cambridge University Press.

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford: Oxford University Press.