Consciousness is a matter of constraint

- 30 November 2011 by Terrence W. Deacon
- New Scientist issue 2840.

A new theory of consciousness depends as much on what isn't there as on what is – and could even help us understand our early origins

IN A 1992 issue of *The Times Literary Supplement*, the philosopher Jerry Fodor famously complained that: "Nobody has the slightest idea how anything material could be conscious. Nobody even knows what it would be like to have the slightest idea about how anything material could be conscious." In 2011, despite two decades of explosive advances in brain research and cognitive science, Fodor's assessment still rings true.

Why is that? Is it just that cognitive neuroscience still has a long way to go? Or have we been looking in the wrong places for clues? For hints to this mystery, brain researchers and philosophers of mind have focused on brain processes, neural computations and their correspondences with the physical world. But what if we should be focusing on what is not there instead?

This proposal is at the heart of my new book *Incomplete Nature*. I believe that in order to overcome this stalemate we need to pay more attention to what is intrinsically not present in everything - from life's functions and meanings to mind's experiences and values.

This suggestion is not intended as an invitation to mysticism, rather it is a way of pointing to the importance of what the field of statistical mechanics calls "constraint": the degrees of freedom not realised in a dynamical process. To illustrate, consider how a quickly flowing stream forms stable eddies as it curls around a boulder, or how a snow crystal spontaneously grows its precise, hexagonally symmetric, yet idiosyncratic branches.

In both cases, the resulting order is a consequence of possibilities that become increasingly improbable by the compounding of constraints, due to continual perturbation. Thus, as the branches of a snow crystal grow, they progressively restrict where new growth can take place. Constraints reflect what is not there, and the more constrained something is, the more symmetric and regular it is.

Notice, too, that the function of an engine, meaning of a word, or content of a thought are also not actually present in the machine, the text, or the firing patterns of neurons. Does this render these missing attributes outside the realm of empirical science? After all, what's not there can't do anything, can it?

Two of the most significant theoretical breakthroughs of modern science may shed some light on this: Darwin's theory of natural selection and Claude Shannon's theory of information. Though separated by nearly a century, both provided powerful conceptual tools that transformed our understanding of the world. Yet despite their familiarity and importance, few recognise the powerful insight unifying these theories: both depend on attending to the relationship between what is present and what is specifically absent.

Darwin showed that selective elimination of that fraction of the excessive variety of organic forms least well-suited to their environment is responsible for the evolution of adaptive functions. Similarly, Shannon demonstrated that information can be precisely measured by comparing a received signal to the variety of signals that could have been received, but were not - thereby reducing uncertainty.

In both cases, what is not present (but could have been) is as important as what is present, whether for determining functional appropriateness or information. Although the possible forms not reproduced during evolution or transmitted in a message are neither material nor energetic, they are still critical influences in

the world. Consider a search team looking for a child lost in the forest. Though 50 people may join in, only one will find the child, but without the other 49, it is unlikely the child would have been found.

Not surprisingly, constraints are also fundamental to the capacity to perform physical work. It is only because there is a constraint on the release of energy - for example, in the one-directional expansion of an exploding gas in a cylinder - that a change of state can be imposed by one system on another.

But can we extend this insight into more mysterious realms where "teleological" causes, those that are purposeful or end-directed, appear to operate? These include the origins of life, the nature of conscious experience, even emotion: all processes that seem to be organised with respect to potentials as yet unrealised. My aim is to provide a thoroughly naturalistic account of how true purposiveness can emerge from purely mechanistic physical processes when they become organised in a way that preserves specific absences, that is, constraints.

Our current scientific predicament reminds me of Zeno's paradox, the Greek riddle where swift Achilles can never overtake a tortoise in a race, or even reach the finishing line, because he must first traverse an infinite number of fractions of that distance. No matter how many details we discover about brains or the quantum fluctuations that might (or not) be taking place inside synapses, we get no closer to a physical account of conscious experience.

Zeno's paradox was solved when mathematicians figured out how to calculate with values that are virtually zero - a trick that ultimately became the basis for calculus. So perhaps this paradox of the mind will only dissolve when we learn how nature operates with the physical analogues of zero - the functions, meanings and experiences by which something virtual may become actual.

Although developments in complexity theory, non-linear dynamics and information theory form much of the background for this theory, something I call "emergent dynamics" theory adds a game-changing twist. It shows how a process I call "teleodynamics" forms a bridge from matter to what matters. In so doing, it leads us into realms where most natural scientists fear to tread.

To explain teleodynamic (end-directed) processes, such as those found in organisms or human minds, we need to step beyond the way complexity and information theories use "constraint", to explain how constraints can become their own causes, how constraints become capable of maintaining and producing themselves. This is essentially what life accomplishes. But to do this, life requires more than self-organisation and more than molecular replication: it must persistently recreate its capacity for self-creation.

Although living systems depend on self-organising processes to create their orderly structures, life must involve more than a mere constellation of self-organising molecular processes because there is no real "self" in self-organisation. What I mean by self is an intrinsic tendency to maintain a distinctive integrity against the ravages of increasing entropy as well as disturbances imposed by the surroundings.

To be truly self-maintaining, a system must contain within it some means to "remember" and regenerate those constraints determining its integrity which would otherwise tend to dissipate spontaneously. This can be achieved when two or more self-organising processes become linked so each one generates the constraints that make up the boundary conditions necessary for the other to occur.

Let's look at this fundamental dynamical transition using a molecular thought experiment involving the linking of two classes of molecular processes. The first involves a small set of molecules forming what is called an autocatalytic chemical process (catalysts that make catalysts so that they reciprocally synthesise each other). The second involves a type of molecule which, in high concentrations, spontaneously tends to self-assemble into hollow containers (like the proteins forming the shells of most viruses). These two molecular processes can become synergistically linked if molecules that tend to self-assemble are produced as side products of an autocatalytic process.

This is because a container will tend to form where catalysts that depend on one another are most concentrated, thus keeping them together, which is necessary for future autocatalysis. The whole complex thus becomes self-maintaining of its own self-maintenance capacity, a process I call "autogenesis".

Such a system tends to reconstitute itself if disrupted, and can even reconstitute multiple replicas of the original if broken up because it reproduces the very constraints that preserve this constraint-preservation process - even with the complete turnover of its constituent molecules. It's the simplest exemplar of an intrinsically end-directed process, whose most fundamental end is maintenance of itself.

At first sight, such a simple molecular system doesn't appear to have much to do with the emergence of function, or value, much less consciousness. But upon closer examination, these teleological-like properties turn out to ultimately depend on some variant of the self-referential circularity of this sort of formative process. This is because the causal circularity between these interdependent, self-organised molecular processes creates an unambiguous site of "self": its intrinsic capacity for self-creation constitutes a precise self/non-self distinction which is independent of any specific material embodiment. And it is this "self" that most needs to be explained before we can even begin to consider the nature of consciousness.

Though it is presented here and in the book as a thought experiment, it is an entirely testable molecular mechanism. Moreover, the evolution of simple, autogenic molecular processes may even have occurred in parts of the outer solar system rich in methane, nitrogen and ammonia, such as on Saturn's moon Titan. In such an environment - lacking in liquid water - complex polymers should tend to form spontaneously and be capable of catalytic and self-assembling functions.

This suggests a multi-phase approach to life in the solar system, where proto-life (simple autogenic complexes) might emerge on outer planets, ride in on comets, get dumped on inner planets where they are exposed to liquid water and, many processes and reactions later, seed life. Learning how to detect autogenesis in a form radically unlike life on Earth might provide new ways to explore the solar system for clues to our origins.

But what about the origin of consciousness? Although the leap from simple autogenesis to subjective consciousness is immense and speculative, this analysis may still provide the first hint to a solution to the dilemma posed by our version of Zeno's paradox. This is because the origins of life and the origins of consciousness both depend on the emergence of self: the organisational core of both is a form of self-creating, self-sustaining, constraint-generating process.

Ultimately, this kind of reciprocal, self-organising logic (but embodied in neural signal dynamics) must form the core of the conscious self. Conceiving of neuronal processes in emergent dynamical terms allows us to reframe many aspects of mental life. It suggests, for example, that the experience of emotion is intimately connected with the role metabolism plays in regulating the self-organising dynamics of the brain's information-generation processes.

This is because self-organised processes are generated by incessantly perturbing a system away from its equilibrium. This essential component of "self" generation is inevitably in tension with the shifting availability of energy in the brain. The metabolic signals we map with fMRI and PET-scan imagery may be serendipitously providing evidence that conscious arousal is not located in any one place, but constantly shifts from region to region with changes in demand.

To explore this in depth, I refer readers to my book. To allay the fears of non-technical readers, however, I must add that the many threads I draw on from philosophy, physics and chemistry to biology and cognitive neuroscience are presented without technical language or mathematical equations. Even the critical insights from thermodynamics, complex systems theory and information theory are explained with familiar examples.

That said, sampling a chapter or jumping to the end in search of the gist will be more confusing than informative. This is because the argument builds step by step to prevent readers falling into the almost irresistible tendency of thinking about these issues using a contemporary, matter-centred paradigm.

While my theory may fall short of a neurologically detailed explanation of how something material can be conscious, and only hints at the extensive work ahead, it might be the first time we glimpse what it would be like to understand the mystery. Could this be the beginnings of a theory of the physical world that doesn't leave it looking absurd that we exist at all?