Going round in Circles

Inman Harvey

Evolutionary and Adaptive Systems Group, University of Sussex, Brighton UK
inmanh@gmail.com

Abstract

Life and cognition are inherently circular dynamical processes, and people have difficulty understanding circular causation. We give case studies illustrating some resulting confusions, and propose that the problems may lie in failing to properly distinguish between similar concepts used to describe both local and global features of a system.

Where to Start?

Artificial Life owes much to Cybernetics, that the influential 1940s/50s Macy conferences referred to as "Circular Causal and Feedback Mechanisms in Biological and Social Systems." McCulloch, conference chair, described (1960) his quest as "what is a number, that a man may know it, and a man, that he may know a number?" Change 'number' to 'thing' and 'man' to 'organism' for the circular core of autopoiesis (Varela et al., 1974): how can organism and its world co-define each other?

Linear causation and explanation is familiar. One starts with a firm foundation of agreed facts, and systematically builds up from there. However *circular* causation is like a Sudoku puzzle where no part of the whole is guaranteed until all the interlocking constraints can be simultaneously satisfied. It is not obvious where and how to start.

Here we show examples of typical traps people fall into when they attempt to understand circular causation. Though aim ultimately at the circularity of full-blown cognition, we deliberately start with minimal examples of what at best might be called proto-agents, If even these cause confusion, how much more treacherous will more realistic cognition be?

Downwind faster than the wind?

Consider the machine of Fig. 1, with but a single component moving part. A boat is constrained to run left or right along a canal. The single shaft shown rotates freely according to the forces transmitted through the air/water propellors. From the density of air and water, propellor details, and drag resistance of the boat, in principle we can calculate what steady-state boat velocity results for a given wind velocity. Our minimal agent has but a single class of behaviour, steady motion, powered by the relative movement between wind and water.

Many readers may doubt that for any parameter settings the boat will move downwind faster than the wind. Youtube videos (usually showing land-based but equivalent versions: search for 'DDWFTTW') have many comments claiming to 'prove' the impossibility. The reasoning usually starts in the most obvious starting place by considering the wind driving the air-propellor (based on the velocity relative to the boat) and thereby driving the boat. If it accelerates up to 10kph, the relative wind velocity drops to zero, surely leaving no further energy available for the boat to reach a higher speed?

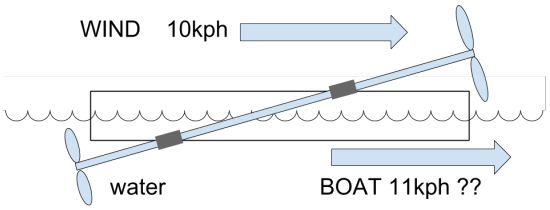

Figure 1: Can it go downwind faster than the wind? Yes.

This reasoning is wrong, despite the mechanism being so simple and the intuitions so compelling. The *actual* steady state solution has the shaft being driven by the water propellor — in the opposite direction to that normally assumed. A linear chain of reasoning starting from the water propellor likewise does not immediately generate this solution, since with an initially stationary boat there is nothing to drive that propellor. The linear reasoning does not fail from starting in the wrong place — there is no right place to start! Circular explanations only work when they include the full circuit of component processes jointly maintaining each other in steady state.

The typical confusion may be in part triggered by using terms such as '*drive*' both locally (wind drives propellor) and globally (wind drives boat) but failing to realise these are different senses. Pre-Copernican astronomy was likewise misled by confusing motion locally relative to Earth with a supposed global motion relative to some universal framework.

How can opposites both be good for you?

Our boat example of circular causation has just one attractor to its dynamics, for a given set of parameters. Our second case study, Daisyworld (Watson and Lovelock, 1983; Harvey, 2015, 2016) has several potential attractors and introduces (at a simplistic level) notions of viability and homeostasis.

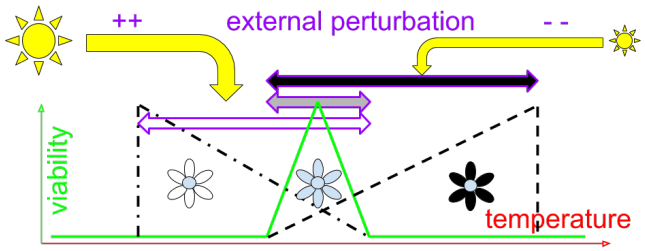

Figure 2: Black, white, grey daisies have same (green) viability dependence on local temp. External forcing from Sun varies. Black increases, white decreases local temp. Feasibility ranges ($\Longleftrightarrow$) of both B&W are extended to lower/higher perturbations.

As summarised in Fig. 2, daisies on a grey planet have a limited range of viability based on their local temperature. A grey daisy derives its temperature from the sun as it alters in solar output (over centuries) and is viable over a limited range of such external perturbations. A black daisy, by absorbing extra heat, extends its range right (to less sun); a white daisy, reflecting heat, will extend its range left (to more sun).

In some sense, then, both increasing (black) and decreasing (white) local temperature is 'good' for the viability of an otherwise neutral daisy. The literature is full of people who disbelieve such a counter-intuitive result, and as discussed in Harvey (2015, 2016) this is largely due to misunderstanding of the circular causation. In particular, there is a tendency to confuse the term '*viability*' that refers to an individual daisy (shown in green on the vertical axis in Fig. 2) with what I now call 'feasibility' referring to the *potential viability* within a range of external perturbations (shown in purple on the horizontal axis); these are (literally) orthogonal concepts.

So again, this may be in part due to a pre-Copernican confusion between local and global concepts.

When do Homeostats or autopoietic entities die?

Our third example of circular causation is the Homeostat (and by extension an autopoietic entity). Ashby's (1952) motivation was to design a machine that learnt through experience to maintain essential variables within bounds. How can a kitten (or machine) learn to avoid the fire without prior knowledge of appropriate input-output responses to heat and pain?

Ashby's answer was to have a mechanism triggered by any crossing of the viability boundary that in essence produced a random variation of the input-output mapping (Fig. 3a). Any inappropriate response would continue further variation, but if and when an appropriate input-output mapping was chanced upon, that formed a stable viable attractor to the circular dynamics. Conceptually this is similar to Darwinian random variation and selection, except within an individual rather than a population; herein lies a problem.

In this context viability is a binary dead-or-alive distinction, a viability boundary is a definite line. But if random variation is only triggered by crossing such a line, surely that is too late, the Homeostat is dead. So the viability signal that Ashby needs for the desired 'ultrastability' is inherently paradoxical.

I believe Ashby made a tactical error, he fell into the same pre-Copernican trap. Viability is *both* a global binary property of an organism, here the Homeostat, *and* a label for a local signal that triggers variation. To identify these global and local senses as the same would be a category error. Here the local sense needs to be analogue not binary (e.g. anything correlated with life-expectancy associated with current essential variables), with that *probabilistically* pulling the trigger for variational change. Even if both 'viabilities' were step-functions, they could not be the same step-function.

This same issue of viability boundary is inherited directly in autopoietic theories (Varela et al., 1974; Fig. 3b), and Di Paolo (2005) makes this explicit. Conservation or breakdown of organisation in an autopoietic system is a binary step-function, so how, Di Paolo asks, can that provide a gradient that confers *significance* of the danger for the organism (e.g. kitten)? The claim appears to be that some such gradient (Di Paolo develops a concept of adaptivity to provide it) is needed for a 'pointer' towards the danger lurking across the viability boundary. But this is to confuse *direction* in essential-variables space with *direction* in the room with the fire. As with *drive* and *viability* in the earlier examples, there is here a mis-identification of local with global meanings of a term.

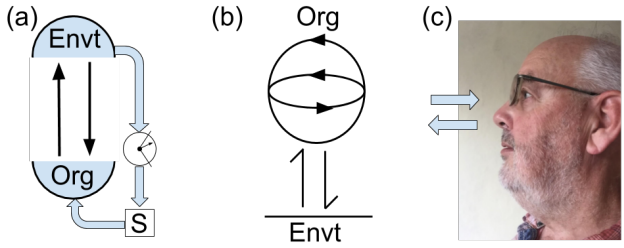

Figure 3: (a) Schematic of Homeostat, 2nd feedback path triggers S if essential variables outside limits. (b) Related schema for autopoiesis, (c) Models a, b are part of my world.

What about us?

Copernicus removed our firm foundation on earth at the centre of the universe. Special relativity eliminated the fixed aether and required new frameworks for physics. Life and cognition are much more complex than physics, and still await their Copernican, relativistic revolutions. Even the simplest proto-agent can confuse us with its unfamiliar circular causation, its lack of a settled foundation on which to build analysis.

It is suggested here that such confusion arises often through mistaking the map for the territory, careless identifying of local and global concepts as the same. GOFAI and even connectionism (Harvey, 1996) is full of such pitfalls, through attributing cognitive agent-level functions to component parts. Sometimes this is useful conscious metaphor. Too often it is taken literally, people fall into the trap; so-called internal representations in the brain would be a classic example.

Autopoiesis takes circular causation most seriously. But even here there is yet further circularity often unrecognised. Physics explains and redefines the 'stuff' of our world, relates tables to atoms. But models of life and cognition are themselves 'stuff' in the world that we live in, we ourselves are inside what we seek to understand (Fig. 3c). The 'stuff' in our models must be different from the 'stuff' of our models.

References

Ashby, W. R. (1952). *Design for a brain.* Chapman and Hall.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences* 4: 429–452.

Harvey, I. (1996). Untimed and misrepresented: connectionism and the computer metaphor. *AISB Quarterly,* 96:20–27.

Harvey, I. (2015). The circular logic of Gaia: fragility and fallacies, regulation and proofs. In Andrews, Caves, Doursat, Hickinbotham, Polack, Stepney, Taylor and Timmis (Eds.), *Proc. Eur. Conf. on Artificial Life 2015*, pages 90–97. MIT Press.

Harvey, I. (2016). Social systems and ecosystems: History matters. In Gershenson, Froese, Siqueiros, Aguilar, Izquierdo and Sayama (Eds.), *Proc. Artificial Life Conf. 2016,* pages 418–425. MIT Press.

McCulloch, W. S. (1960). What is a number, that a man may know it, and a man, that he may know a number? *General Semantics Bulletin*, 26/27:7–18.

Varela, F. J., Maturana, H. R. and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems* 5: 187–196.