*Artificial Life*, 2019, Accepted, In Press

# Circular Causation, Circular Cognition: a Tour around some Common Confusions

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. We analyse how explanations in terms of circular causation work and how they rely on principles of 'Normal Settlement'. Even though they typically will not explain the *origins* of phenomena (that is the province of linear causal explanation), circular explanations have predictive power for any *persisting* (i.e. stable or metastable) phenomena.

Keywords: Causality, Complex Systems, Dynamical Systems, Epistemology, Feedback,

# 1.0 Where to Start?

Artificial Life owes much to Cybernetics. The influential 1940s/50s Macy conferences referred to cybernetics as “Circular Causal and Feedback Mechanisms in Biological and Social Systems.” McCulloch, conference chair, described [29] 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 [37]: how can organism and its world co-define each other? Circular causation is central to understanding cognition, whether biological or artificial.

*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 we aim ultimately at the circularity of full-blown cognition, we start with a minimal non-cognitive example of circular causation, DDWFTTW (‘Direct Down Wind Faster Than The Wind’). Despite its minimality, many find this challenging to understand. We use this, together with an even simpler example of a natural rock arch, to illustrate the basic concepts of circular causation.

Then we progress through a number of further examples extending into the cognitive domain. These are used to illustrate how explanations based on circular causation work, what principles they are based on, what limitations they have. We identify some patterns in the common errors made.

It is astonishing how little is taught about circular causation despite its centrality to so many important fields (especially biology, cognitive science and artificial life) and despite the abundance of confusion and fallacies arising from widespread unfamiliarity with the concept. Areas of study with the strongest track records for discussing circular causation would include Cybernetics [41] and autopoiesis, enaction [37, 38, 39]; but even these do not focus on the particular confusions and misunderstandings we address here. Hence this paper presents analytic studies with a didactic motivation; in my opinion there is a need for such education. It is extended from an ECAL conference paper [20] and the text and illustrations borrow freely from there.

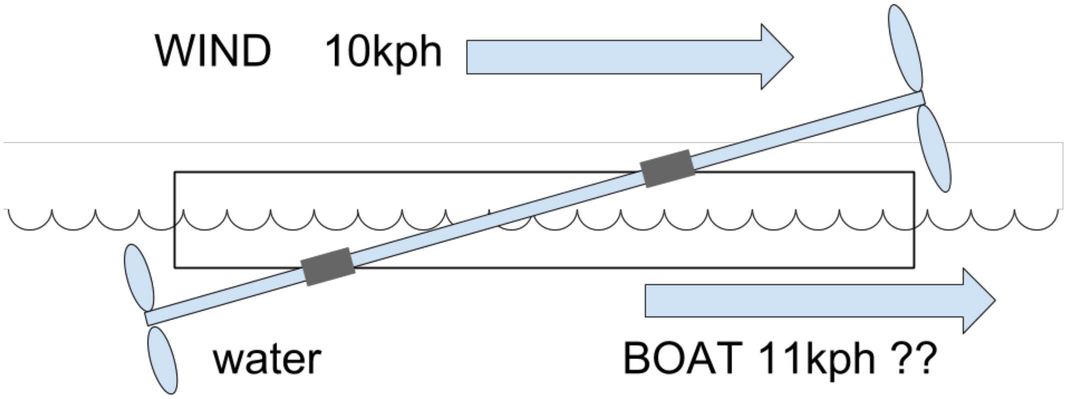

Figure 1: Can it go downwind faster than the wind? Yes.

## 1.1 DDWFTTW: Direct downwind faster than the wind?

Consider the machine of Figure 1, with but a single component moving part. A boat is constrained to run left or right along a canal. The single shaft shown is positioned in a sleeve and 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. We are told that the director of the Lawrence Berkeley National Laboratory refused to allow an external speaker to present a talk on this on the grounds that such a vehicle would violate the laws of physics[1]. 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 force available to propel the boat to a higher speed?

This reasoning is wrong, despite the mechanism being so simple and the intuitions so compelling. The *actual* steady state solution has the shaft rotating in the opposite direction to that normally assumed. I.e., if the handedness of the air propellor was such that wind from the left would drive an unconstrained version clockwise, in fact its coupling with the water propellor results in it rotating anti-clockwise. 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 in translation to linear reasoning when they include the full circuit of component processes jointly maintaining each other in steady state.

For a full analysis one has to calculate the torque on each propellor (at steady-state these torques sum to zero); and the linear forces on each propellor (that at steady-state sum to a total that equals the drag of the main hull). Formally, we have boat speed B and shaft rotation R governed by:

dB/dt is an increasing function of ($\text{Thrust}_{\text{water}}$ + $\text{Thrust}_{\text{air}}$ + Drag) (eqn 1)

dR/dt is an increasing function of ($\text{Torque}_{\text{water}}$ + $\text{Torque}_{\text{air}}$ ) (eqn 2)

when all the terms are themselves functions of B, R and propellor properties such as thrust coefficients, torque coefficients, advance ratios, efficiency, size, handedness. The details are too complex to explore here. But it can be shown [16] that a whole range of parameter choices allow this apparently paradoxical boat motion 'downwind faster than the wind' — just as different parameter choices will allow boat motion upwind.

Navigators of sailing boats will be more familiar with these issues than most. As well as being familiar with the ability to tack into the wind (progressing through zig-zags at around 45° to the oncoming wind), they will know that the fastest way to sail downwind is to tack on a broad reach at around 45° to the

[1] For mention of this episode within an entertaining journalistic account of how Rick Cavallaro and colleagues had such difficulties with disbelievers, see the article in Wired Magazine, 27 August 2010. https://www.wired.com/2010/08/ddwfttw/

downwind direction. This tactic is used in modern high-speed racing, particularly with ice-yachts [8]. Despite the extra distance involved in such tacking, the speed made good in the downwind direction can outpace the wind by a significant factor. If one transplanted such a speeding ice-yacht, on a downwind reach (or upwind tack) on a flat expanse of ice, to some imagined large ice-tube with the same wind blowing down (or up) it — we conveniently here ignore gravity — such an ice-yacht would spiral down the tube faster than the wind itself (or up the tube against the wind). The sails of the yacht would act much as the blade of a propellor — and the circular motion of this blade translates into linear motion downwind (or upwind). Combine such possibilities with the two propellors of Figure 1, and with the energy source from differential motion of air and water, then 'downwind faster than the wind' becomes easily achievable [16].

Why do so many people find this paradoxical and hard to believe? Firstly because of unfamiliarity with the circular causal reasoning required, in terms of the attractors of steady-state dynamics. But a second related reason may be the easy ambiguity of using terms such as '*drive*' both locally (wind drives propellor) and globally (wind drives boat) but failing to realise these are different senses. This is a version of the mereological fallacy [7], or more generally a basic category error. Pre-Copernican astronomy was likewise misled by confusing motion locally relative to Earth with a supposed global motion relative to some universal framework.

## 1.2 Linear causation, chickens and eggs

We assume here that the motivation for wanting an explanation is to be able to predict events, or in other words to avoid surprise. The default form for an explanation for a phenomenon is in terms of linear causation: "We accept without question premises A and B, and we can then deduce consequence or effect C; subsequently, from A B and C we can deduce D, and thence E, F… in linear sequence." This is the way we build a tower out of bricks in successive stages on a firm foundation, or a mathematical proof derived from axioms. For each phenomenon, for each brick in the tower, for each lemma in the proof, we can provide lines of reasoning or of support back to firm foundations. An annoying young child, repeatedly asking "Why?" to each proffered new level of explanation, is appealing to this form of causal explanation. They may be seeking the foundational assumptions, that ultimately can only be given as "Because!"; or they may just be being annoying.

Such forms of explanation are so common that we usually omit the 'linear' qualification, and assume that all causal explanations will take this form. Hence the perceived need for an answer to: "Which came first, the chicken or the egg?" Explanations of origins must necessarily be linear through time, must appeal to linear causation. But in contrast to this we here start to move towards systems of circular causation, where a number of components are mutually interdependent. This puzzles some, as we shall see.

## 1.3 Irreducible complexity and rock arches

Creationist advocates of 'intelligent design' focus on what they call the 'irreducible complexity' of biology where there is their version of circular causation: ".. a single system which is composed of several well-matched, interacting parts that contribute to the basic function, and where the removal of any one of the parts causes the system to effectively cease functioning" [6]. Their issue is not so much

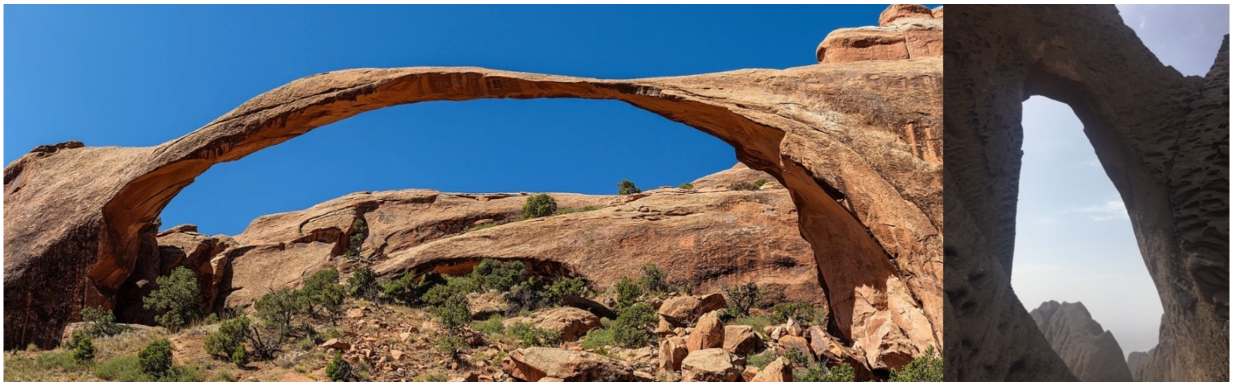

Figure 2. Irreducible simplicity? Natural rock arches: the world's widest (Landscape Arch, Utah, 88m wide: Thomas Wolf www.foto-tw.de, Wikimedia Commons, CC BY-SA 3.0) and the world's tallest (Shipton's Arch, Xinjiang, 460m tall: IH).

with the *functioning* or *persistence* of such systems as with their *origin.* If the parts are mutually interdependent, how could they have been assembled incrementally through successive viable stages, such as implied by a Darwinian explanation?

The answer is illustrated by natural rock arches (Figure 2), that though simple nevertheless fit the definition of irreducible complexity[2]. There are mutually dependent component parts (e.g. for a simple choice of segmentation, left and right halves L and R) that do indeed mutually support each other; remove either one and the other shifts in consequence, probably with the collapse of the arch. There is a further component, the underlying ground G, that is also necessary to provide support; but G is independent of L and R since the removal of either leaves G unchanged. The natural processes that formed the arch did not require the hand of a creator during assembly; successive stages of addition (deposits of sediment) followed by subtraction of 'scaffolding' (erosion) are sufficient. Darwinian evolution is similarly not constrained by the paucity of imagination of the creationists, and is quite capable of assembling through successive viable stages of subtraction, neutral change and exaptation as well as addition. The fact that the maintenance of an end product may be explained through circular causation is no barrier to a separate linear causal explanation for its origin.

## 2.0 Circular causation as used in explanations

The arch provides an intuitive picture of circular causation; let's formalise this. In particular, we characterise when and how explanations of a phenomenon can be framed in terms of circular causation, and what are the limitations.

We use explanations to reframe complex phenomena or patterns, possibly difficult to understand, in terms of simpler entities or components that we do indeed comprehend. We can build an explanation on the known effect that each component A, B, C…Z has on one or more if the others. If such causal chains feed back on themselves in circuits A→B→…→A, then we have the potential for circular explanations.

[2] Cairns-Smith [9] used the same example of an arch for discussion of very similar issues. I thank two reviewers for pointing this out.

If such circuits can be broken down into independent sub-circuits (e.g. A⇿B and C⇿D), then they can be analysed independently. Hence it is sufficient to just consider sets (A…Z) where each member is directly or indirectly potentially affected by every other member.

The rock arch is a minimal example, where each pier (L and R) can take on one of two values: e.g. 1 for standing and 0 for collapsed. There are two possible steady stable states for {L,R}: {1,1} for a standing arch, and {0,0} for complete collapse. We are assuming that the lean of each pier is such that neither can stand unsupported so, though we can imagine an instantaneous snapshot taken of situation {0,1} or {1,0}, neither is a viable long term proposition.

Generalising from this, we can say that a circular explanation involves explaining the existence of (one or more) attractor(s) in a system of several interacting components with circular chains of cause and effect; all components, both individually and jointly, need to be in stable equilibrium. Such an equilibrium may be a static point attractor, as with the arch, or a steady state of dynamic equilibrium, as with DDWFTTW, or indeed any other form of cyclic or even chaotic attractor. Such an explanation is useful even where we may not have a linear historical explanation of origins available; we can still predict how external perturbations will shift or collapse such a system. Such explanations inevitably have no obvious starting point, and we shall see below how this can be confusing.

## 2.1 Timescales and Circular causation

Any physical system that displays patterns significant enough for us observers to contemplate and discuss must inevitably have some associated rough range of timescales. Circular causation, being at root based on the concept of attractors, is no different. For natural rock arches the range might lie roughly from a few seconds to many years. An unsupported pier might take a few seconds to collapse, so one needs a longer period of observation than that to be happy that an arch is stable. At the other end of the scale, the fact that all such arches will have changed or disappeared after millennia does not stop us calling them currently stable over many years.

Typically we ignore events faster than the fast end of this timescale, except in so far as they will destabilise any unstable equilibrium. We may call this ‘noise’, that ensures the dynamics heads into attractor basins. But at the top of the range we may often choose a variety of timescales for consideration, when we consider metastability and the difference between ‘variables’ and ‘parameters’. On a timescale of years, rock erosion can probably be ignored and we can treat a rock pillar or arch as stable and fixed. When however we shift our perspective to that of millennia, erosion becomes a variable that we must consider significant. A ‘parameter’ typically describes an effect that counts as a variable on a long timescale, but can be considered as fixed over a shorter timescale. It is often necessary, when considering explanations in terms of circular causation, to make explicit such shifts in timescale perspective. This may become clearer in the Daisyworld example below, where solar insolation is considered as a fixed parameter for the purposes of temperatures settling into an attractor — but nevertheless may vary over longer timescales.

## 2.2 Sequential and steady-state causation

In linear causation, where A has an effect on B that has a knock-on effect on C…, effects follow causes sequentially in time. At first sight this sequential property seems incompatible with circular causation.

But not really, since circular causation is based on steady-state analysis, and sequential cause-and-effect is crucial to understanding whether a circuit A→B→…→A in steady-state is stable or unstable. The reasoning involves considering a random small shift to one component, e.g. A, when the whole circuit is in some specific steady-state situation, and then following the consequential chain of cause and effect around the circuit until it reaches A again. If a small positive (or vice versa: negative) shift at A results in a consequent negative (or vice versa: positive) response on A, the circuit is in negative feedback and hence stable. With the reverse response, it is unstable. So the analysis in terms of stable and unstable steady states that circular causation usually relies on is at a higher conceptual level than linear cause-and-effect alone, but nevertheless builds upon such sequential linear chains; no incompatibility.

It might be suggested that A→B→A should be rewritten as something like A(t)→B(t+1)→A(t+2), for the update ticks of some universal time t. I should emphasise that this would be a highly undesirable misinterpretation. It confuses the map with the territory, confuses a computational simulation of a world (that typically uses clock ticks) with the real world (that does not). The reference in the previous paragraph to the consequences of 'a small shift' in one variable should be read as 'an arbitrarily small shift, over an arbitrarily small time'. Mathematically speaking, we learn in calculus to define dx/dt as the limit, as $\delta t \rightarrow 0$, of $\delta x / \delta t$, where x takes value $(x + \delta x)$ at time $(t + \delta t)$. Stability of x at time t is defined as dx/dt=0 at time t (and not at any time t+1).

Why does this apparently insignificant distinction raise any concerns? Because the failure to appreciate this has led to too many simulations, in Artificial Life and elsewhere, failing to correctly model their target worlds through inappropriate choice of update time-steps. This typically results in computational artefacts in the simulation, such as numerical instabilities, that mislead; we shall give examples in the Daisyworld section 4.2 below. One pragmatic way to ensure that the simulation choice of update time-step $\delta t$ is small enough is to make a first guess of an appropriate $\delta t$ and run the relevant part of the simulation, recording the result; then repeat with a $\delta t$ that is 1/10th the size, then again 1/10th smaller … until the results do not materially change in value. This corresponds closely to the literal interpretation of defining dx/dt as 'the limit, as $\delta t \rightarrow 0$, of $\delta x / \delta t$'. Simulations are invalid if there has been no check as to whether an inappropriate choice of $\delta t$ has been made; at minimum a basic 'reality check' is needed, and ideally this more principled check[3] using decreasing values for $\delta t$.

# 3.0 Circularity taken to extremes: holograms

The simplest forms of circular causality involve a small number of mutually dependent variables, as in the mutually supporting piers of a stone arch, or the dynamical relations of a boat with propellors. More complex circularity can arise with multiple interconnected units forming multiple overlapping circuits. Thomas and colleagues [33,34,35] present interesting perspectives on these, particularly in the biological context of cell differentiation. What happens when you take this towards the extremes of millions of tiny interacting units, approaching an analogue continuum?

One version of such an extreme is illustrated by a hologram [15], Figure 3. The viewers looking through the holographic film see an image behind in 3D, that offers them the same shifts in perspective

[3] Though the Matlab examples shown in this paper do not explicitly show such principled reality checks in the code, they were indeed done.

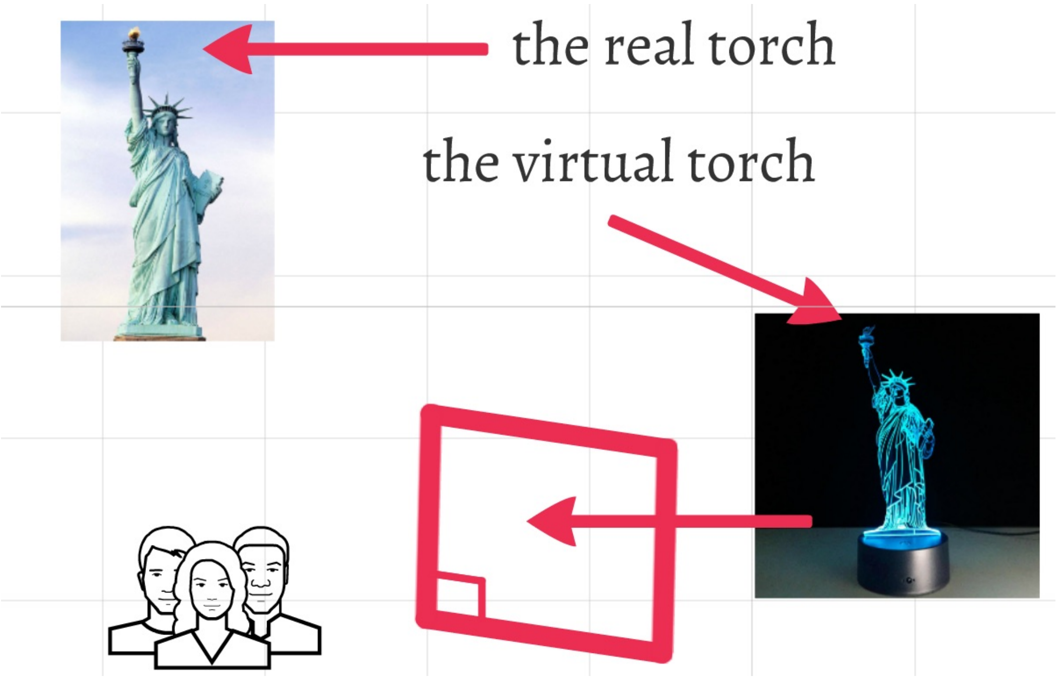

Figure 3. The virtual statue appears to the viewers to be in 3D behind the holographic film, and image parts, such as torch, have the same relationship to the whole as in the real statue. But there is no direct 1-to-1 correspondence between any small segment of the holographic film and some segment of the image.

as they move around as the shifts when they are viewing the original. Whereas in normal photography different parts of film record different parts of the image, a crucial feature of holograms is that the whole of the image is recorded everywhere across the film. So if e.g. only a small part of the film (shown at lower left in Figure 3) is available, then the whole of the image can be viewed (at somewhat lower resolution) by peering through that segment at an appropriate angle.

This property means that the holographic film has the useful feature of graceful degradation, that loss of parts of the film does not lose any part of the image. The whole is still available, though the image quality starts to degrade. As such, the hologram model has been very influential as a source of inspiration for how brains may facilitate the storage of memories [14,22]. The metaphor suggests that just as with holograms there is no one-to-one mapping between image features and locations on the celluloid film, there is no need for any one-to-one mapping between individual memories and any locations in the brain that might be 'storing those memories'. These ideas lie at the root of recent deep learning advances in AI; Hinton, recipient of the 2018 Turing Award for his foundational contributions to deep learning, was influenced in his original research directions by exactly this hologram metaphor [14].

In what sense do holograms involve circular causation? Appropriate laser light focussed on the interference patterns recorded on the celluloid film results in the creation of a light field, or optic array, that is in crucial aspects identical to the light coming from the original object. It is instructive (and somewhat counter-intuitive) to identify what is happening by reference to Figure 3, and the role of the observers depicted there. These observers will agree on the 3D location of the virtual image; if they each point at it, those pointer lines will intersect at that 3D location. Yet if we go round ‘behind the film’, nothing is there, we can pass our hand through empty space. There are not even any relevant light waves emanating from that specific location. In the absence of the observers *there is nothing there!*

However in the presence of those observers, the reconstructed wavefronts [15] resulting from the constructive and destructive summation of light waves from everywhere on the available image film will mutually support each other to form stable perceptions by those observers; perceptions that are interpreted by them as if there was an object there. The agreement by the observers as to the location of the virtual image remains stable even as the observers shift their viewpoints, even as parts of the celluloid film are removed or corrupted.

The stability, in location and form, of the virtual image (despite noise on/partial removal of the celluloid film, despite observer movement) follows the same pattern as the stability in location and form of the stone arch (despite any ground tremors). But whereas the stone arch is composed of real stone at its actual location, the virtual holographic image is not composed of anything material at its location, not even from light waves arising from there. The virtual image arises from the collaboration between (a) the observers with their visual perception attuned to normal visual circumstances and (b) the optic array from the celluloid film illuminated by laser. Just as left and right pillars of a stone arch support each other, here (a) observers and (b) that optic array support each other to maintain a stable virtual image at a location where nothing really exists!

Stability of something as sophisticated as the perceived form and location of a virtual image in the eyes of a human observer might seem far distant from the stability of an arch. But they are both examples of stable equilibria within circular causation A→B→A. Look for the consequences of noise, look for invariants despite such noise. In the above sketch of the hologram the only noises considered so far have been disruption of the celluloid film and observer movement, as illustration of the principle; a fuller analysis would include further consideration of the human perceptual system.

## 4.0 Moving towards agent-based circular causation

Our first two examples, DDWFTTW and the arch, were inanimate, as is the hologram. We now move towards more animate models of circularity. Living creatures both depend on the world around them, and have effects on that world, and hence entail circular feedback loops.

We start with basic environmental interactions, in a simple yet still counter-intuitive artificial life model. From there our examples progress to more full-blown examples of cognition.

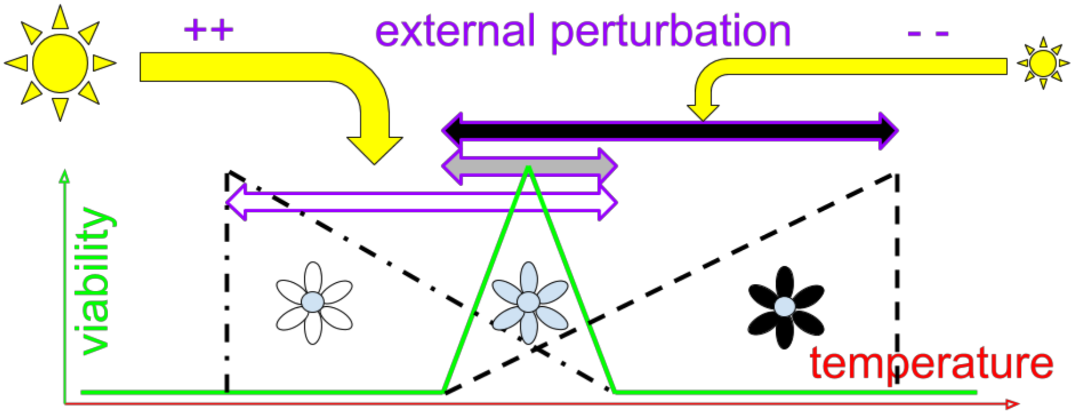

Figure 4: Black, white, grey daisies have same (green) viability dependence on local temp. External forcing from Sun varies. Black increases, white decreases local temp. Feasibility ranges (⟺) of both B&W are extended to lower/higher perturbations.

## 4.1 Daisyworld: how can opposite effects 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 next case study, Daisyworld [41,18,19] has several potential attractors and introduces (at a simplistic level) notions of viability and homeostasis.

As summarised in Figure 4, the model assumes that 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 potential range left (to more sun); the extension of range in one direction is not penalised by a corresponding diminution in the other direction.

The core result of Daisyworld Theory, as I have argued elsewhere [18,19], can be summarised with a 'single-Daisy' version thus:

$$\frac{dD}{dt} = H(T) - D \qquad \text{(eqn 3)}$$

For a current local temperature T, the daisy population D moves towards some value specified by a non-negative viability function H(T). Let *Viability Range* S be the support of H, in mathematical terms the range of T values for which H(T) is non-zero; i.e the range of temperatures for which daisies are viable.

$$\frac{dT}{dt} = P + kD - T \qquad \text{(eqn 4)}$$

For some externally fixed perturbation P (e.g. derived from current solar output) and the current daisy population D, the temperature move towards a value influenced by the quantity of daisies D upwards if $k > O$ (black daisies absorb heat) or downwards if $k < 0$ (white daisies reflect heat). This coupled pair of equations in T and D will have one or more stable equilibrium solutions with D either zero ('extinct') or positive ('extant'). For any given values of k and P we define $Q_k(P)$ as the maximum D value at any such stable equilibrium. The *Feasibility Range* of k, F(k) is the support of $Q_k(P)$; the range of P values for which there is at least one stable equilibrium solution with positive D. We can show [18,19] that when k=0, F(0)=S the support of H; though F(0) is a *Feasibility* range of P-values and S a *Viability* range of T-values, here they happen to be measured in the same units (though in more complex

Daisyworlds they will not). But the important Daisyworld result is that for all non-zero k, $F(k) \supseteq F(0)$ and there is always some finite 'big enough' k=K (whether positive or negative) such that $F(K) \supset F(0)$. In simple terms, make k big enough (+/-) and you will increase the feasibility range; whatever k, it is impossible to reduce the feasibility range below that for k=0.

For a simple demonstration we may use a simple hat-shaped viability function:

$$H(T) = \max(0,1 - |2T|) \qquad \text{(eqn 5)}$$

This is a 'witches hat' shape, with a peak value at T=0 decreasing to zero for T anywhere outside the *viability range* [-0.5, 0.5]. We give here[4] basic Matlab code for numerically simulating the above equations (eqn3, eqn4, eqn5). Though this is the simplest example with just one daisy-type, it serves as an introduction to the core principles before advancing to more complex Daisyworlds with two daisy-types or more.

In this sense of promoting increased feasibility range, 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; some of the most prominent are cited and discussed by Harvey [18,19], pointing out that 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 Figure 4) with what is here called '*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.

## 4.2 Daisyworld and Numerical Instabilities

Simulations of Daisyworld can mislead if they are not constructed so as to pass the reality check discussed above in the section 2.2 on Sequential and steady-state causation. A classic example comes from Kirchner, a Gaian sceptic. In a review of Daisyworld scenarios [26], he described one (his Figure 3 and associated text) that exhibited "pathological" behaviour, an unstable oscillation between only-black and only-white daisies. My simulation version of this scenario came to a stable (and non-pathological) equilibrium with intermediate values, as expected. In an attempt to understand how our simulations differed, from correspondence with Kirschner (personal communications, May 2015) I established to my

[4] Basic demo (just 1 daisy-type) Daisyworld Matlab code:

```
delta=0.01; kc=0; data=zeros(41,41);
for k=-4:0.2:4                  % step through k values, both - and +
   kc=kc+1; pc=0;               % counters
   for P=-2:0.1:2               % step through P values
      T=0; D=0; pc=pc+1;        % initialise each time
      for t=1:100000            % iterate to stability
         D=D+delta*(max(0,1-2*abs(T))-D);      % update D
         T=T+delta*(P+k*D-T);                  % update T
      end                       % should be stable
      data(kc,pc)=D>0.00001;    % note when D is viable (use near-zero as cutoff …
   end                          % … to account for numerical approximations)
end        spy(data);           % show Feasibility ranges for P, at different k-values
```

satisfaction (if not to his) that this was a classic case of the simulation time-step failing a basic reality check.

If

$$\frac{dB}{dt} = (W - B) \qquad \frac{dW}{dt} = (B - W) \qquad \text{(eqns 6)}$$

is simulated by Matlab code such as:

```
delta=0.01; B=1;W=0;
for t=1:10000
   Bdelta=delta*(W-B);
   Wdelta=delta*(B-W);
   B=B+Bdelta;
   W=W+Wdelta;
end
```

with a suitably small value for the time-step *delta*, then [B,W] settles down from initial values [1,0] towards equilibrium at [0.5,0.5]. But if *delta* is set to a value 1, the behaviour does indeed oscillate "pathologically" between [1,0] and [0,1] . The former non-pathological version is the correct simulation for such differential equations, the latter with *delta* = 1 is wholly misleading.

Advocates of Daisyworld as well as sceptics are similarly vulnerable to such errors. Even Lovelock's most cited paper introducing Daisyworld [41], whilst qualitatively reasonably correct, nevertheless contained significant quantitative errors for related reasons, as pointed out in [24].

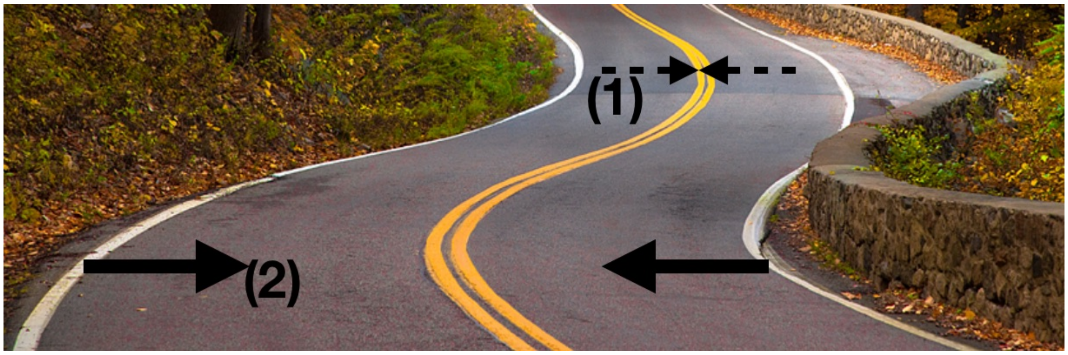

Figure 5. Different strategies for staying on the road: (1) move to centre line (set-point). (2) stay away from edge.

## 4.3 Rein control: stability without a set-point

The Daisyworld phenomenon is a form of homeostasis since it involves stabilising some essential variable (e.g. room temperature or body temperature) at or around some appropriate value. There are basically two possible classes of mechanism for achieving this, illustrated in Figure 5.

The first — often wrongly assumed to be the only way — is the way central heating systems regulate room temperature *T* despite varying external temperature $T_e$. We need the regulatory system to provide a correcting temperature offset $T_c$ such that $T = T_e + T_c$ settles down to the desired $T_{setpoint}$. The difference between between actual temperature *T* and the set-point desired temperature is measured; depending on whether this is positive or negative, the system acts so as to reverse this.

$$\frac{dT_c}{dt} = M(T_{setpoint} - T) \qquad \text{(eqn 7)}$$

Here the function *M( )* is a monotonically increasing function, in its simplest form a linear function $M(x) = kx$ with $k>0$; here the size of *k* sets the timescale of the response to any perturbations. The system temperature would then move to $T = T_{setpoint}$ regardless of what value the external perturbation $T_e$ takes. More realistically, such a biological system will have limits beyond which the system fails to efficiently regulate. Nevertheless, the hallmark of this traditional class of feedback system is: within such limits there is regulation to a *fixed set-point*.

The second biological style of homeostasis is different and has no set-point; engineers typically are unaware of this[5]. Clynes [11] was among the first to point out that biological control systems operate with internal control variables that can only take non-negative values — and hence in order to control against both being ‘too hot’ and ‘too cold’ they must have two separate pathways for doing so. Sometimes called the principle of unidirectional rate sensitivity, Clynes also called this rein control, a metaphor that is easier to understand. Since the reins of a horse can only pull, not push, in order to steer a horse in either direction you need different reins on each side — and biological control systems do just this.

The rein control version is to use two separate control variables, pulling against each other so as to reach an equilibrium; a classic biological example would be antagonistic pairs of muscles. We illustrate here with an example rein control system for staying away from both edges of the road (Figure 5.2) to reach equilibrium somewhere in between. Let x represent the position on the road relative to the centre line at x=0; positive x is on the right of the road. There is an external perturbing force, say the wind W, tending to push a vehicle (or horse) to one side or the other. Our rein control consists of one or two correcting reins, $C_L$ and $C_R$ (Figure 6). We use the same naming convention as horses’ reins, with right rein $C_R$ being the one responding to being too far to the left (negative x); hence $C_R$ is only zero or positive, $C_L$ is only zero or negative.

The effect on x of these perturbations and correcting forces may be summarised as:

[5] A search of IEEE publications for ‘rein control’, via IEEE Xplore Digital Library, finds just one mention, namely a paper by the present author. Google Scholar offers 57 citations for Clynes’ main paper on this [11], none of them from a (non-biological) engineering context.

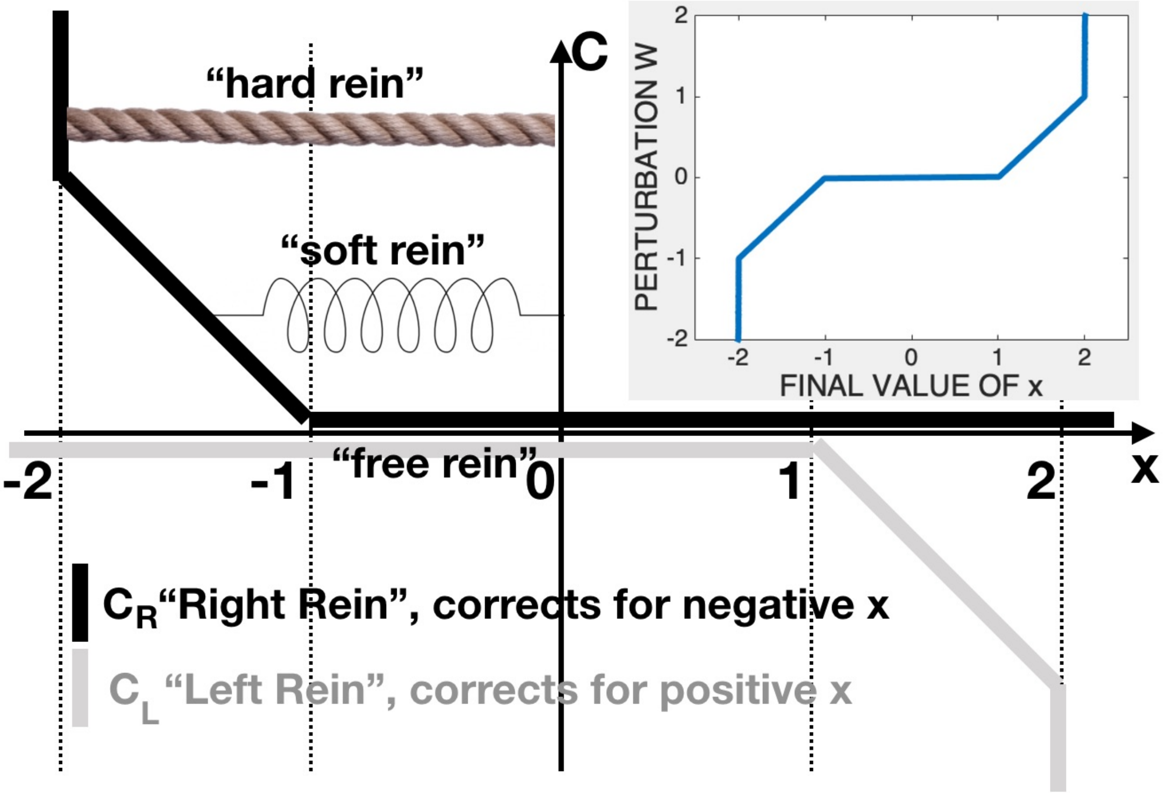

Figure 6. The heavy black line illustrates one possible example of a Right Rein correctional force $C_R$. The equilibrium value of $C_R$ is: 0, for $x > -1$ ("free rein"); $1 - x$, for $-2 \leq x \leq -1$ ("soft rein" indicated by spring); and 'as large as is needed', for $x < -2$ ("hard rein" indicated by rope over-riding spring). $C_R$ will be restrained from going to infinity by the circular causation of x being dependent on C, as well as C dependent on x: see equations in main text. Heavy grey line illustrates a similar possible Left Rein $C_L$. The inset plot shows final x-values for a rein control system using such reins $C_L$, $C_R$. to correct against different perturbations W; see main text.

$$\frac{1}{\tau}\frac{dx}{dt} = (W + C_L + C_R) - x \qquad \text{(eqn 8)}$$

with an equilibrium at $x = W + C_L + C_R$ .

Different positive values for $\tau$ in the various differential equations will set different timescales for each. It would be important to account for this if fully analysing the dynamics, but for our present purposes of equilibrium analysis we may simplify by assuming all $\tau = 1$.

The right rein $C_R$, as illustrated in Figure 6, obeys:

$$\frac{dC_R}{dt} = \max(0, -1 - x) - C_R \quad ; \text{if } x > -2 \qquad \text{(eqn 9a)}$$

$$= (-2 - x) \quad ; \text{if } x \leq -2. \qquad \text{(eqn 9b)}$$

The left rein $C_L$, as shown, behaves similarly:

$$\frac{dC_L}{dt} = \min(0,1-x) - C_L \qquad ; \text{if } x < 2 \qquad \text{(eqn 10a)}$$

$$= (2-x) \qquad ; \text{if } x \geq 2. \qquad \text{(eqn 10b)}$$

We append Matlab code for simulating these equations (eqn 8, eqn 9a/b, eqn 10a/b) numerically[6]. The resulting plot (inset in Figure 6) shows that x is maintained within the 'hard boundaries' [-2,2], regardless of W values. But there is no set-point, the final value of x is dependent on W. This example is just one from a large family of possible 'reins' with any combination of "hard" or "soft" or "free". Conventional negative feedback control to a set-point can be seen as a specific restricted member of this family where both $C_L$ and $C_R$ have a "hard rein" fixed to that one set-point.

## 4.4 Why call this homeostasis?

What, then, justifies calling this a form of homeostasis, when the resulting equilibrium value of $x$ will vary according to the perturbing external force $W$? Because as $W$ varies across a range of values, equilibrium $x$ varies across a *reduced* range, less than in the absence of such reins. This can be seen from simple mathematical considerations, and (as an aside for chemists or economists familiar with this) it would appear to be a very basic example of Le Châtelier's Principle [27] in action. This Principle is usually stated as something like 'the typical consequence of any perturbation to a system in equilibrium is a shift to a new equilibrium that partly counteracts the perturbation'. In many cases, such as Daisyworld, this 'counteraction to a perturbation' results in viability being possible over a wider range of external perturbations than would otherwise have been the case. Even if one does not stay at the centre line, one still stays on the road.

We note that Clynes' rationale [11] for biological systems use of such strategies was the observation that internal bodily control signals used variables, such as chemical concentrations, or frequency/intensity of spikes, that could vary in positive values but not go below zero. Though artificial systems may not be so constrained, as an aside for those familiar with Artificial Neural Networks it may be of interest that this relates directly to the ReLU or Rectified Linear Unit activation function $f(x) = \max(0,x)$ that has recently become popular. This basic nonlinearity is surprisingly powerful.

[6] Basic demo rein control Matlab code:

```
delta=0.001; c=0; data=zeros(401,2);
for W=-2:0.01:2                          % step through possible W values
   x=0; CR=0; CL=0; c=c+1;               % initialise
   for t=1:1000000                       % settle to equilibrium
      if x> -2    CR=CR+delta*(max(0,-1-x)-CR);     % soft or free rein
      else        CR=CR+delta*(-2-x);               % hard rein
      end                                           % Right Rein updated
      if x< 2     CL=CL+delta*(min(0,1-x)-CL);      % soft or free rein
      else        CL=CL+delta*(2-x);                % hard rein
      end                                           % Left Rein updated
      x=x+delta*(W+CR+CL);               % x updated
   end                                   % should be stable
   data(c,1)=W; data(c,2)=x;             % store data
end       plot(data(:,2),data(:,1));     % show final x-values for each W
```

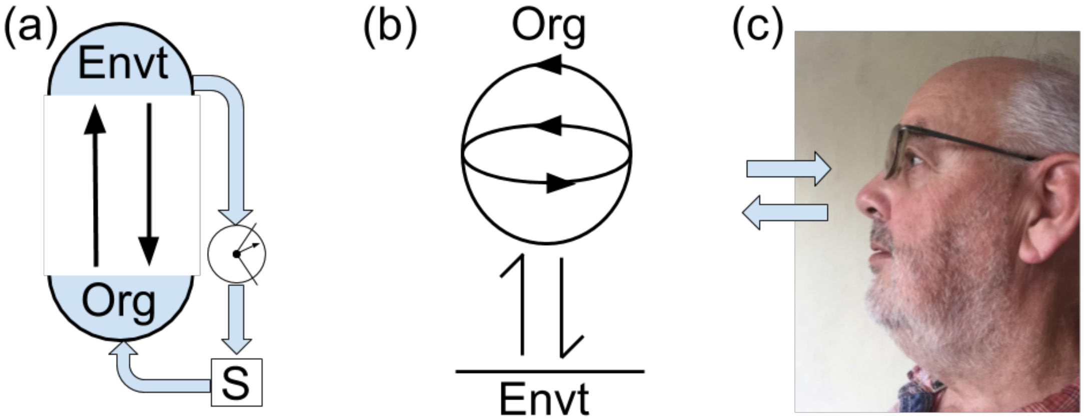

Figure 7: (a) Schematic of Homeostat, 2nd feedback path triggers S if essential variables outside limits. (b) Related schema for autopoiesis, (c) Models a, b are part of my world.

# 5.0 Learning adaptive behaviour

Our next example of circular causation is the Homeostat (and by extension an autopoietic entity). Ashby's [2] 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 caused some random variation of the input-output mapping (Figure 7a). 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.

## 5.1 When do Homeostats or autopoietic entities die?

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 [37], see Figure 7b, and Di Paolo [13] 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 it is a category error to identify *direction* in essential-variables space with *direction* in the room with the fire; as with *drive* and *viability* in the earlier examples, this is a mis-identification of local with global meanings of a term. There should be some correlation between the two, but necessarily there cannot be identity between them.

One obvious class of candidate local signal would be any scalar entity that at least crudely negatively correlated with life-expectancy. For a kitten approaching a fire, then heat and (when getting too close) associated pain has just that crude negative correlation. With kittens, we assume that some such crude correlation has been promoted through Darwinian natural selection. But this is not the only explanation for such correlations — even natural rock arches subject to erosion (that clearly do not have origins in Darwinian evolution) tend to react to local rock falls (crudely negatively correlated with arch life-expectancy) by rearrangement of forces in a manner (Le Châtelier's Principle again, see Section 4.4) that promotes their survival. How? — through 'Normal Settlement'.

# 6.0 Circular causation and Normal Settlement

This paper has attempted to show why circularly causal systems are often so baffling. The type of explanatory strategy proposed is based on what I like to dignify as Harvey's Universal Principles of Normal Settlement (NS) — the pompous title no doubt failing to conceal that these are just restatements of the blindingly obvious:

(NS1) Any complex system will spend more time in some parts of its phase space than others; the more visited including attractors of all sorts, stable and (over extended periods) metastable states.

(NS2) If you find the system in one of the common regions, you should be less surprised than when you find it in a less common region!

If you agree that a defining factor (or, as I would claim, *the* defining factor) for an explanation is that it should reduce the degree of surprise in the observer by underwriting (implicit or explicit) provisional predictions, then it follows from NS principles that identifying the more common regions of phase space is a step in the right direction. Sometimes this step is bigger than others; if the system can be shown to have a single point attractor (and sufficient time to reach it) this is a stronger result than where there are multiple attractors with little evidence to distinguish between them.

Explanations presented on this basis are frequently by some standards incomplete. For example they may fail to address the origins of a phenomenon, even that as simple as the stone arch example above. But nevertheless they can be predictive, and it is clear how they can be modified and improved. whenever a new phenomenon surprises. Incidentally, this style of explanation lies at the root of Darwinian Natural Selection, a different NS that is a specialisation of the more universal NS principles presented here. Natural Selection = Normal Settlement + Heredity.

## 6.1 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 is full of such pitfalls [17], 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 (Figure 7c).

Models of biological or cognitive systems, as used in biology or Artificial Life, are typically conceived from an assumed 'God's-eye' viewpoint outside the system of study. We need to carefully distinguish between what is perceptually available from our privileged external observer viewpoint and what is perceptually available to the organisms or agents of study. A satirical novel (and early Alife speculation) that highlights this issue is Flatland [1,21].

This describes a plane 2D world in which people are simple geometrical polygons or lines, and their world and interactions with it follow the consequent 2D constraints. Whereas we, from our 3D perspective, can imagine a Sphere passing through the plane, to its inhabitants this (or its intersection with their world) would appear as a dot out of nowhere, expanding into a circle of maximum diameter, that then dwindles and disappears again. In turn, these 2D inhabitants are aware of further constraints that hold for a 1D 'Lineland' or even a 0D 'Pointland'. Visualising lesser dimensions is relatively easy, but for them visualising their position within higher dimensions is too much of a challenge.

This novel, though now dated, nevertheless highlights a core issue when we try to understand other worlds and our own world. If we choose to assume (whether wisely or not) that ultimately the agents in that world and the objects they interact with are basically made of the same 'stuff', then consequences follow. For agents in that world, objects arise (or are perceived, or enacted) out of stuff-stuff interactions — and this cannot be the same as stuff. So as a matter of logic, what is 'matter' to an agent in that world cannot be the same as any 'stuff' that both agents and objects are built from. Failure to make this distinction is the same kind of category error as the confusions between *local* and *global* meanings of a term highlighted in earlier sections of this paper.

This is circularity taken to the limit, but we need to face up to the fact that cognition is not just a phenomenon that inescapably involves circular causation — it is a circle that includes us ourselves within it (Figure 7c). The 'stuff' in our models must be different from the 'stuff' of our models.

## 6.2 Comparisons with other types of explanation

Conventionally the background section of a paper, citing other relevant work, appears at the beginning. By citing existing work in the field, we present the foundation on which new results can be built, in the tradition of linear causal explanations. For this paper it seems more appropriate to place it here at the end, for two reasons. Firstly, there is surprisingly little previous work on circular causation to cite; we have already mentioned that Cybernetics [41] and autopoiesis, enaction [37, 38, 39], whilst discussing circular causation, do not focus on the particular confusions and misunderstandings we address here. Secondly and more importantly, it is only after developing the analysis above that we can compare and contrast with existing ideas that look related — but on inspection may turn out to be somewhat different.

The view of circular causality presented here clearly overlaps with discussions on dynamical systems approaches to cognition (e.g. [4,36]), and with the approach associated with earlier work in Evolutionary Robotics (ER) (e,g. [5,23]). The importance of attractors in such analysis was sometimes explicit [23]. Ashby [2] was an important early influence. It is commonplace within ER to view the task of the roboticist as that of shaping the behaviour space of robots so as to have the desired attractors.

Some would relate the view presented here to notions of 'emergence'. I reject that label, since it has been so widely used in such vague undefined terms as to be unhelpful. Where attempts have been made to define emergent explanations clearly [3], they have been negative definitions: emergence has been defined in terms of filling a gap in the observers' expectations — this corresponds to my understanding of explanations as a response to surprise — but then there is rather little specified as just *how* this gap is being filled.

The view presented here should be seen as incompatible with some attempts [10] to subscribe to a dynamical systems or enactive approach to cognition that nevertheless retains cognitivist and representationalist explanations. This confuses representations as *explananda* (crucial features of our human cognition that need explaining) with representations as *explanans* (used as part of an explanation). It is an example of a basic category mistake confusing brains with minds or persons — while the view presented here is trying to make such distinctions very clear. The cognitivist carelessly elides the *global* sense of 'representations', as used by people in everyday life, with a *local* sense as used metaphorically by a homunculus in the brain [17].

There is an ongoing debate in the philosophy of science literature [25,30] about the difference between (i) 'mechanistic' explanations in biology and (ii) explanations that use 'dynamical and mathematical models'. On first reading, there could be loose parallels here with what I have been calling explanations via (a) 'linear causation' and (b) 'circular causation'. But Silberstein and Chemero [32] see 'localization and decomposition' of the component parts of explanations as a prime distinguishing factor between (i) (that has 'localization and decomposition') and (ii) (that is not characterised thus). The arguments presented here distinguishing (a) linear from (b) circular explanations do not appeal to this factor.

There is an interesting connection between the ideas of circular causation presented here and repeated attempts (e.g. [12,28]) to use notions of 'information' as some crucial ingredient that explains why or how living organisms apparently cheat the 2nd law of thermodynamics. Schrödinger [30] brought ideas of information into biology, specifically genetics, and the tools of information theory have been invaluable. But technical usefulness has been too often mixed with analytical carelessness, with category errors. It looks to me that these ideas rely on Shannon information, and as Shannon [31] makes clear, his theory would have been better named communication theory than information theory. Shannon information is specifically a relational concept, relative to a sender and receiver communicating in the presence of noise. This places it squarely within the type of context in which circular causal explanations are appropriate, and principles of Normal Settlement are crucial. It also means one needs to distinguish carefully between *local* (information relative to component parts of a system conceived as metaphorical homunculi) and *global* (including information from an external 'God's-eye' perspective) senses of information — and this is where too many fail to do this and fall into exactly the same trap as the cognitivist's misuse of representations discussed above. In summary, these mistakes are symptoms of the proponents of information in this context failing to understand how explanations based on circular causation work.

# 7.0 Conclusions

Of course circles do not have conclusions, but papers do, and we should summarise the main points to stress.

We define circular causation as what we appeal to in causal explanations of phenomena that rely on multiple contemporaneous interactions between component parts. Whereas linear explanations are built up incrementally, with each stage in principle fully explained and settled before the next stage is added, in contrast circularly causal explanations require the full circle complete before they work.

When we refer to causality (circular or linear), we here really refer to the type of explanation (circular or linear) that we use. The role of an explanation is taken as allowing predictions and reducing surprise. 'Circular *reasoning*' typically refers to a logical fallacy, and this may in part explain why the concept of circular *causation* is often treated with suspicion. But mainly people are just unfamiliar with the idea, remarkably little is written about it.

Circularly causal explanations typically appeal to principles of 'Normal Settlement'. We identify regions of phase space of a system that are more likely (often overwhelmingly so) to be visited than the average. So often we are dealing with dynamical systems, and their attractors and metastable regions. We typically explain why *if* a system is in a specific state or in a specific attractor *then* it is likely to stay like that. Typically we may not be offering any explanation for *why* the system arrived in that state in the first place. Linear explanations are usually needed for origins.

Circularly causal phenomena provoke lots of confusion in people unfamiliar with them. Examples such as DDWFTTW and Daisyworld have many disbelievers, with many proofs offered that the phenomena cannot be real. Though convincing to some, these 'proofs' are flawed.

One signature error that often accompanies this confusion is a misidentification of terms used *locally* to describe some component part of the system, with similar terms used *globally* to describe the global behaviour of the system. Examples given above include the concept of *drive* in DDWFTTW, of *viability* (and synonyms) in Daisyworld [18], of *direction* in Di Paolo's analysis of autopoiesis [13]. In neuroscience such confusion is common, and Bennett and Hacker [7] refer to this as the mereological fallacy. I prefer to call it a straightforward category error, that circular causation seems particularly prone to. The cognitivist confusion between *representations* that people use in their life and 'representations in the brain' is a notorious example, and I have suggested above that *information* is likewise misinterpreted by some.

Biology and cognition — and thus necessarily Artificial Life — are domains that are permeated throughout by multiple overlapping feedback circuits, both within organisms and further through their coupling with the environment. It is incumbent on us to recognise the underlying principles needed for explanations, to be aware that our intuitions (too often based on linear assumptions) may be wrong, and to look out for the typical errors identified here. Ultimately, if we want to understand human cognition, we must be prepared for the consequences of being within the cognitive circle ourselves, not just outside observers.

## References

[1] Abbott, E. A. (1884). *Flatland: a romance of many dimensions.* Seeley and Co., London.

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

[3]Baas, N. A. and Emmeche, C. (1997). On emergence and explanation. *Intellectica,* 2(25):67-83.

[4] Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Science* 4(3): 91-99.

[5] Beer, R., and Gallagher, J. C. 1992. Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1(1), 91–122

[6] Behe, M. (1996). *Darwin's black box*. New York: The Free Press.

[7] Bennett, M. R. and Hacker, P. M. S. (2003). *Philosophical foundations of neuroscience.* Blackwell Publishing.

[8] Bethwaite, F. (2010). *Faster sailing techniques*. Adlard Coles Nautical.

[9] Cairns-Smith, A. G. (1985). *Seven Clues to the Origin of Life*. Cambridge University Press,

[10] Clark, A. (1998). *Being there: putting brain, body, and world together again.* MIT Press.

[11] Clynes, M. (1969). Cybernetic implications of rein control in perceptual and conceptual organization. *Ann. NY Acad. Sci.* 156:629-670.

[12] Davies, P. (2019). *The demon in the machine*. Allen Lane.

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

[14] Ford, M. (2018). *Architects of intelligence.* Packt Books.

[15] Gabor, D. (1949). Microscopy by reconstructed wavefronts. *Proc. Roy. Soc. A,* 7 July 1949:454-487.

[16] Gaunaa, M., Øye, S., and Mikkelsen, R. (2009). Theory and design of flow driven vehicles using rotors for energy conversion. *Proc. EWEC 2009*, Marseille.

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

[18] 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.

[19] 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.

[20] Harvey, I. (2017). Going round in circles. In Carole Knibbe et al, (Eds.) *ECAL2017 Proc. Eur. Conf. Artificial Life 2017*, pages 198-199. MIT Press.

[21] Hinton, C. H. (1907). *An episode on Flatland: or how a plane folk discovered the third dimension*. Swan Sonnenschein, London.

[22] Hinton, G.E. and Anderson, J. A. (1981). *Parallel models of associative memory.* Erlbaum.

[23] Husbands, P., Harvey, I. and Cliff, D. (1995). Circle in the round: state space attractors for evolved sighted robots. *Robotics and Autonomous Systems* 15:83-106.

[24] Isakari, S. M. and Somerville, R. C. J. (1989). Accurate numerical solutions for Daisyworld. *Tellus* 41B, 478-482

[25] Kaplan, D., and W. Bechtel. 2011. Dynamical models: an alternative or complement to mechanistic explanations? *Topics in Cognitive Science* 3:438–44.

[26] Kirchner, J.W. (1989). The Gaia hypothesis: can it be tested? *Reviews of Geophysics* 27:223-235.

[27] Le Châtelier, H. (1884). Sur un enoncé général des lois des équilibres chimiques. C*omptes Rendus,* 99:786-789.

[28] Maynard Smith, J. and Szathmáry, E, (1995). *The major transitions of evolution.* Oxford University Press.

[29] 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.

[30] Schrödinger, E, (1944). *What is life? The physical aspect of the living cell.* Cambridge University Press.

[31] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal.* **27** (4): 623–66

[32] Silberstein, M. and Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philosophy of Science,* 80:5 (Dec 2013) 958-970.

[33] Thomas R. and D'Ari, R. (1990). *Biological Feedback*. CRC Press, Boca Raton, Florida, USA.

[34] Thomas, R. L. and Kaufman, R. (2001a). Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11(1): 170– 179.

[35] Thomas, R. L. and Kaufman, R. (2001b). Multistationarity, the basis of cell differentiation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. *Chaos*, 11(1):180–195.

[36] Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21(5):615-628.

[37] 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.

[38] Varela, F. J. (1984). The creative circle: Sketches on the natural history of circularity. In P. Watzlawick (Ed.), The Invented Reality (pp. 309-324). New York, NY: W. W. Norton & Company, Inc.

[39] Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: MIT Press.

[41] von Foerster, H. (1979). *Cybernetics of Cybernetics.* In K. Krippendorff (Ed.), Communication and Control (pp. 5-8). New York, NY: Gordon and Breach.

[41] Watson, A. J. and Lovelock, J. E. (1983). Biological homeostasis of the global environment: the parable of Daisyworld. *Tellus* 35B:284-289.