Appears as: *Artificial Life* 24(1):29-48 (2018)

Robustness and Contingent History: From Prisoner's Dilemma to Gaia Theory

Inman Harvey

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

Abstract

In both social systems and ecosystems there is a need to resolve potential conflicts between the interests of individuals and the collective interest of the community. The collective interests need to survive the turbulent dynamics of social and ecological interactions. To see how different systems with different sets of interactions have varying degrees of robustness, we need to look at their different contingent histories. We analyse abstract Artificial Life models of such systems, and note that some prominent examples rely on explicitly a-historical frameworks; we point out where analyses that ignore a contingent historical context can be fatally flawed. The mathematical foundations of Gaia Theory are presented in a form whose very basic and general assumptions point to wide applicability across complex dynamical systems. This highlights surprising connections between robustness and accumulated contingent happenstance, regardless of whether Darwinian evolution is or is not implicated. Real life studies highlight the role of history, and Artificial Life studies should do likewise.

Introduction

In both ecosystems and social systems there are at least two levels at which, speaking loosely, 'lifelike' processes can be observed. There is one level at which the individual organisms, animals, humans are interacting with each other and pursuing their individual interests. But also there is a second level of ecosystem organisation, or social organisation, which provides the context within which they exist. In principle the same individuals could function, perhaps more or less successfully, if the ecosystem/social organisation was changed. One extreme version of such a change would be for the ecosystem/social organisation to break down in chaos, which is often against the interests of the individuals concerned. Systems survive or die, much as individuals do.

Ecosystems and Social Systems. Some social organisation can be the outcome of a social contract where individuals have chosen to agree to a set of rules. Other social organisation, and all ecosystems, do not involve such explicit choice. Regardless of such differences, in both cases one can analyse individual behaviour in terms of self-interest potentially clashing with the interests of others around. In social systems we may call some actions 'cheating' and some consequences 'punishment'. In ecosystems we tend to avoid such moral overtones and merely discuss 'effects' and 'consequences'; the analyses may nevertheless be similar.

How do they Persist? If a specific ecosystem/social system survives for a long time, explanation is called for. If no external authority is responsible for imposing this, then the organisation must be an emergent consequence of individual patterns of behaviour that are globally somewhat resilient to the perturbations of everyday life. We may ask how *one specific* ecosystem/social system manages to persist, or we may ask about *generic* properties needed for persistence.

What is their Origin? Each specific ecosystem/social system will have its own unique history, from origins up to the present day; just as each organism has its unique genetic and developmental history. It is the main thesis of this paper that *generic* theories, that gloss over or average such *specific* histories, often fail to capture salient features of reality. Examples of such theories will be criticised.

Real Life. We consider both real systems and their artificial life counterparts, as in different columns of Table 1. For real social systems we look at common pool resource governance as studied by economists. Natural systems refer here to ecosystems as studied by ecologists in the field. It will be suggested that those studying such real life systems will have no problems agreeing with the thesis that history matters. Hence this paper is mainly targeted at those producing abstract models that explicitly leave out any consideration of history.

Artificial Life Models. Simple abstract models of social systems are illustrated here by examples from IPD, Iterated Prisoner's Dilemma. This is based on a classic two-person game where each player has simple choices and the interactions between them have consequences in terms of different payoffs. The basic dilemma of individual cheating versus cooperation is distilled into this simplest form. Ecosystem models discussed here are based on Daisyworld models where the organisms and environmental

	Real Life	**Artificial Life**
Social systems	Common pool resources [18]	Iterated Prisoner's Dilemma [19, 21]
Natural systems	Ecosystems Niche construction [3, 13]	Daisyworld [23, 7] Complex systems [17]

Table 1: Classes of decentralised social systems and natural (eco-)systems and their Alife counterparts analysed here.

influences are characterised as variables interacting in a dynamical system. Robustness and the role of contingent history are considered in the context of new results in Gaia Theory.

Mathematical Summary. If processes are actually non-Markovian, modelling them as Markovian will lead to error.

Plain Language Summary. Real life takes place in a world of accumulated historical accidents that affect how social and ecological processes actually function. History matters.

The contingent history of this paper is that it is an extension of Harvey [8] as presented at ALife XV in Mexico, 2016. Significant portions of text are thus taken directly from that paper with minimal editing, so that this paper may be read without need to consult its immediate precursor. What is notably new here is the latest version of the Gaian Regulation Theorem, summarised in the text and laid out in full in Appendix B. Key to this is a newly elaborated central concept of *feasibility*, related to but importantly different from viability.

Artificial Life Models of Governance

The Introduction to Hobbes' Leviathan [9] gives the first known reference to artificial life under that name:

> NATURE (the art whereby God hath made and governes the world) is by the art of man, as in many other things, so in this also imitated, that it can make an Artificial Animal. For seeing life is but a motion of Limbs, the beginning whereof is in some principall part within, why may we not say that all Automata (Engines that move themselves by springs and wheels as doth a watch) have an artificiall life?

This introduces the metaphor of a nation state as an artificial man, Leviathan, with different components functioning together as mechanically deterministic parts but forming a living whole. What sort of governance can provide some form of global harmony ensuring cooperation and collaboration between component parts and reconciling any potential conflicts between them? How does some form of social contract arise (and continue to survive) from a natural state of anarchy? Hobbes' answer was for central rule by an absolute sovereign. Though such a sovereign is ultimately driven by his private interests, these are aligned with the public interests in so far as "The riches, power, and honour of a monarch arise only from the riches, strength, and reputation of his subjects".

Here we follow Hobbes' surprisingly modern notion of artificial life, and the use of models such as Automata in the study of real life governance. However we part company with him on his assumption of a need for a central sovereign. Leviathan is Hobbes' exemplar of central authority, but in all further examples we discuss below there is no central authority, no rules for behaviour are imposed from outside.

Distributed Social Systems, Choice, Attractors

When governance arises solely through interactions between individual participants, different styles of governance can only arise through the different choices they make. Strategies any one individual has for choosing are typically conditional in part on the choices the others have made. These individual choices bind into a social system when there is a stable pattern that persists despite potential disturbances from within or without. In dynamical systems terms, we are looking for the attractors of such systems. There may be many possible such attractors, some more congenial than others to the participants — e.g. with higher payoffs in utilitarian terms.

Ecosystems, Choice, Attractors

With natural ecosystems we may not be considering the same type of explicit strategies or choices seen in social systems. Nevertheless, a different type of 'choice' is available, a choice of where to locate, which environment to inhabit. Animals may move from valley to hilltop; even plants, over generations, can shift their habitat. In this subtly different sense of choice, the component members of an ecosystem have 'chosen' to coexist in a specific locale where their various interactions (including their own knock-on effects on the environment) allow them to thrive. In the theoretical space of all conceivable ecosystems, there is a multitude of such viable and robust locales that act as potential attractors.

Real Social Systems: Common Pool Resources

Real social systems with distributed governance are only stable if they are currently near an attractor, which is another way of saying that they recover from small disturbances. We look at how such systems may adapt to changing circumstances over the longer term, and hence the role of historical contingency in how they come to be in one attractor rather than another. Practical working examples of decentralised control can be seen in societies across the world ranging from water authorities in California to shared forest usage in Nepal and Switzerland and shared fisheries in Turkey. These are maintained and policed by the participants themselves rather than imposed by some external sovereign authority. Such 'common pool resources' have been the focus of economist Ostrom [18]. She proposes a list of design principles or 'best practices' that are common to such robust institutions:

(1) Clear identified boundaries between those people and resources *within* the institution and those *outside*.
(2) Appropriation rules congruent to local social and environmental conditions.
(3) All (or most) members share in making or changing rules.
(4) People who are users (or accountable to them) monitor the appropriation and resource management.
(5) Sanctions for rule violations are graduated from low to high according to the severity or persistence of violations.
(6) Conflict-resolution mechanisms are local and rapid.
(7) External authority, e.g. higher government, does not enforce its own rule contrary to that of the local institution.
(8) Where there are multiple levels of governance they are organised in multiple nested layers.

In such common-pool scenarios, anonymous entry or participation is not possible. All participants have not only a stake in *maintaining* the rules (principles 4 and 6) but also in *changing* them (principle 3). Such adaptation in the governance system needs to be congruent with local social and environmental conditions (principle 2); and the social conditions may include further higher or lower level layers of governance, overlapping in a nested fashion (principle 8). Within the generic constraints of these 8 principles there is scope for a multitude of possible governance systems each adapted, more or less, to local circumstances and fashioned through a historical succession of contingencies.

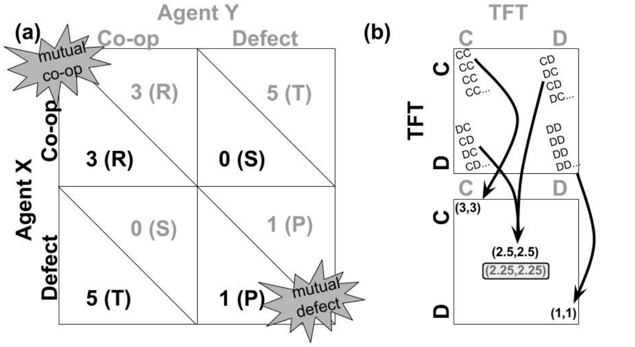

Figure 1: (a) The IPD payoff table [19]. (b) Tit-for-Tat players differentiate into TFT_C or TFT_D (opening play C or D). Different varieties meeting (upper square) lead to 3 different end-attractors, average scores of (3,3), (1,1) and (2.5,2.5) (i.e. average of (5,0) and (0,5), lower square). The weighted (25%, 25%, 50%) average of all these attractor scores is different yet again, (2.25,2.25).

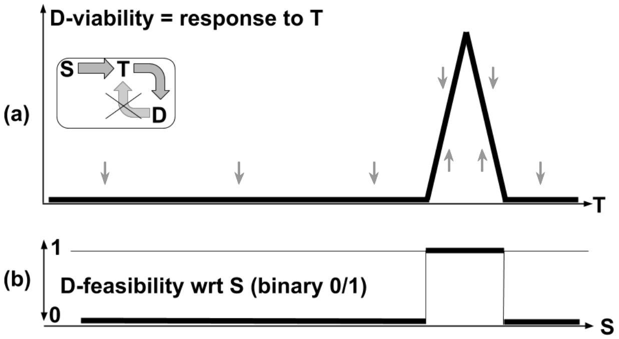

Figure 2: No-feedback scenario: environmental perturbation S (solar output) directly affects local environmental T (temperature) which directly affects organism D (daisies). (a) D assumed to have steady-state dependency, 'hat-shaped' function of T, giving limited zone of viability. (b) D-feasibility (binary yes/no) plotted against perturbation S

ALife Models of Social Systems: IPD

We move from stability, contingency, history in *real* systems to the same issues in ALife models. Recent innovations in IPD (Iterated Prisoner's Dilemma) provide a case study.

Motivation for IPD Models. These provide a minimal model of 2 agents ('prisoners') interacting. They must decide on actions independently, but the payoff to each depends on what they both decide, and is designed to provide a conflict between individual and collective gains.

The supposed story is that they have agreed beforehand to deny everything about some joint crime, but now they are interviewed separately by the police. Each has to decide whether to keep quiet as promised ('Cooperate' with the other prisoner) or make some deal with the police ('Defect'). In terms of utility, they both receive R (say 3) if both Cooperate; both receive P (1) if both Defect; and if one Defects, the other Cooperates the payout is T (5) to the defector and S (0) to the other. The choice of (T, R, P, S) = (5, 3, 1, 0) (Figure 1a, following [19]) meets the IPD condition T>R>P>S that implies whatever agent 2's decision is, agent 1 would gain

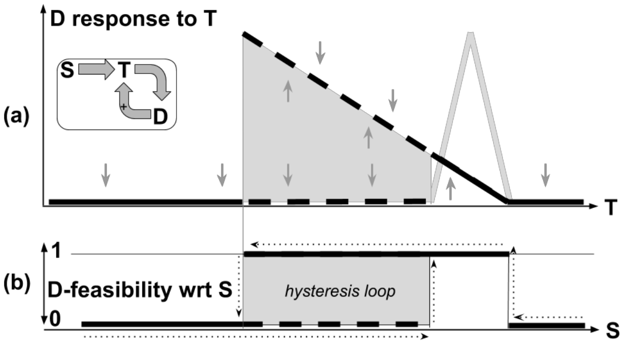

Figure 3: As Fig. 2 plus a further influence of D on T (here positive, black daisies increase temperature) (a) Peak response of D to changes in S is shifted, with a hysteresis loop. (b) D-feasibility zone is extended (here to the left) by a buffer zone, only effective if entered from high S values, and not low ones.

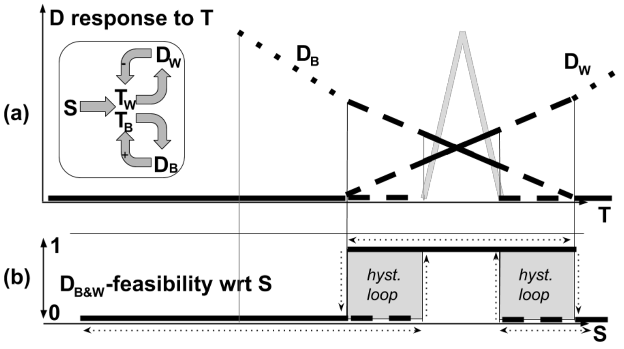

Figure 4: As Fig. 3 plus white daisies D_W affect negatively their local temperature T_W as well as black daisies D_B affecting positively T_B. (a) shows steady-state values of each D (b) shows feasibility of D_B&D_W (simultaneously), against S.

more by Defecting than Cooperating. The further condition 2R>T+S implies the total payout for both Cooperating, 2R, is higher than the total payment when one Cooperates and the other Defects.

The rules treat each agent symmetrically, so any difference in outcome depends solely on how their strategies interact. In a single game with no further consequences, each agent maximises their payoff by Defecting, irrespective of the other agent's choice. Hence they both Defect (D), receiving 1 each, whereas if both Cooperated (C) each would have received 3.

If such games are iterated indefinitely, in the IPD, then each agent's actions may influence future responses. Under some circumstances a regime of Cooperation for mutual benefit can arise; IPD studies usually focus on just what conditions allow this and discourage cheats (i.e. Defectors). Such conditions provide counter-examples to Hobbes' intuition that only a sovereign authority can guarantee a mutual Commonwealth.

Tit for Tat, TFT. A typical class of IPD strategy depends (either deterministically or probabilistically) on memory of the previous choices made by each agent in the previous N rounds, N≥1. Tit for Tat (TFT), for example, is the memory-1 strategy where an agent copies the action that the other agent took in the previous round ([1]; Figure 1b). Tit-for-Two-Tats is the memory-2 strategy where an agent only defects if the opponent defects twice in a row. More generally a memory-N strategy can be specified as a table with 4^N cells, relating to 4 possibilities CC, CD, DC, DD (for own+opponent choices) on N previous rounds, each cell specifying the probability that C will be chosen by this agent in the new round. For example, TFT with memory 1 has these probabilities of Cooperating, dependent on the previous round: CC 100%, CD 0%, DC 100%, DD 0%.

Historical and A-Historical Agents

Such memory-1 strategies depend explicitly on short-term memory of the previous round; but they also depend on long-term history of starting conditions, since the very first move makes a difference, say C for TFT$_C$ or D for TFT$_D$. There are two possible routes to finesse this issue, the first being to acknowledge that TFT$_C$ and TFT$_D$ are indeed two different strategies with different consequences (Figure 1b). Only when TFT$_C$ meets another TFT$_C$ does the virtual circle of Cooperation take off. If both are instead TFT$_D$ then a vicious circle of Defection takes over. A TFT$_C$ meeting TFT$_D$ results in alternating CD, DC choices. The starting conditions have a permanent effect on which basin of attraction is entered.

A second way to finesse this issue is to arrange affairs so that initial conditions eventually become irrelevant, and this could be the case with sufficient noise in the system. If with high enough probability a choice is accidentally reversed, then over enough iterations of IPD all possible basins of attraction will be visited. In a classic early Alife paper [15] Lindgren explicitly used this method. There is a cost to be paid for finessing matters this way, however: TFT$_C$ and TFT$_D$ are now indistinguishable in such a theory, despite the fact that over any finite run they typically have totally different behaviours.

In principle the IPD game iterates for an arbitrary number of rounds, not known in advance. If both players know it to be the final game, this becomes a one-shot PD where both must rationally Defect. In turn, the penultimate game falls to the same analysis, and so on back to the first. An infinite series of rounds avoids this trap, but is impossible in practice. But we can have a finite, non-predetermined, number of rounds by arranging *after* each iteration a small (e.g. 1%) chance that it is *then* deemed to be the last. Then if any noise (as introduced by Lindgren) is small in comparison to this 1%, strategies such as TFT$_C$ and TFT$_D$ will be visibly seen to operate in different basins of attraction. Real world scenarios typically resemble this pattern rather than the infinite-iteration limit. For such real world scenarios, the history will matter.

This distinction between historical and a-historical agents is the central focus of this paper. Behaviour of the latter depends only on recent short-term events held in 'memory', whereas the former also depends on one-off longterm origins in history. Crudely, this can be related to different perspectives from Biology and Physics. Typically many biologists are interested in a specific species or in ecosystems with a specific evolutionary history (which we can relate to TFT$_C$ or TFT$_D$). In contrast physicists, broadly speaking, may be happier making broad generalisations across some arbitrary range of entities (which we can relate to Lindgren's TFT); often this makes the mathematics more tractable. Taken to extremes, this can result in broad statements that are generically true about "all possible organisms” assuming ergodicity, thus including extant organisms on this planet together with all extinct organisms, and indeed all conceivable organisms on all conceivable planets; but nevertheless misleading about any one specific non-ergodic organism. What is true about generic a-historical IPD agent TFT can be false about TFT_C or TFT_D.

Press and Dyson

A recent ground-breaking IPD paper by Press and Dyson [19], displayed a novel class of memory-1 ZD (Zero Determinant) strategies. These allow an agent — provided it no longer had the simple ambition to maximise its own payout that traditionally is expected in IPD — to tailor its strategy to guarantee that the opponent’s payout will average some value such as 1.5 (between P and R) regardless of how the opponent responds. Or such an extorting agent can guarantee that the excess of payoff above P will be shared in unequal proportions such as 3:1. The details of these ZD strategies are not discussed here. They are highly novel and counter-intuitive and are acknowledged by others to be valid, given the context; but many of the conclusions Press and Dyson drew have been shown to be misplaced [22]. We summarise these points, then go even further in questioning the validity of their Markovian assumptions..

Extortionate ZD Strategies. Suppose agent_X chooses an extortionate ZD strategy that gains a bigger proportion of the excess rewards (above a base-level of P) regardless of agent_Y’s responses. Then if agent_Y is an optimising player that adjusts strategy so as to increase its own payoff (Press and Dyson call this an evolutionary player) the result is that agent_X scores even higher. The erroneous implication Press and Dyson draw is that in an evolutionary scenario where multiple strategies are competing against each other, such extortionate strategies will triumph and dominate. As Stewart and Plotkin [22] and other commentators point out, this is not so. If extortionate players came to dominate an evolutionary scenario, they will typically be competing with similar extortionate strategies. If agent_X and agent_Y are both forcing their excess payout (above P=1) to be 3 times greater than their opponents, this is neatly resolved by the excess being 0 for each, the (1,1) score of mutual Defection.

Generous ZD Strategies. It turns out that so-called Generous ZD strategies — that roughly speaking do the opposite of extortion in making sure that differential benefits mostly accrue to their opponents — will dominate in an evolutionary scenario. Such Generous strategies behave optimally against other Generous strategies, and also replace non-cooperative ZD strategies [22].

Such ZD Strategies Ignore Historical Contingency

The main contribution of this paper to this novel development in IPD studies is to point out what other commentators have apparently missed: this whole class of ZD strategies, whether extortionate *or* generous, has been set up to be a-historical and hence to be largely irrelevant as models of human (or animal) strategies — since these human strategies are typically historical, contingent and contextual. Press and Dyson [19] explicitly set up their ZD strategies to use the same finesse Lindgren [15] uses, as discussed above, to average over all possible contingent longterm histories; they focus on generic strategies dependent on short-term memory alone. Indeed, they go further than Lindgren in showing that such Markovian assumptions allow any memory-N strategy to be generically equivalent to (some other) minimal memory-1 strategy.

Their proof covers the TFT strategy averaged over all possible histories, but fails to cover a TFT_C strategy, even with its short history of a single first move. *A fortiori*, such IPD results have even less relevance to the real world when e.g. analysing the mating behaviour of *this* specific butterfly, with its long evolutionary and ecological history of multiple over-lapping constraints as context; or when analysing the governance system for *that* Turkish communal fishing arrangement, with its long social and cultural history of multiple over-lapping polycentric social contracts. Ostrom [18] explicitly mentions congruence with local social and environmental conditions among her design principles observed in long-lasting common pool governance systems, and this historically contingent context is what is stripped away in such generic proofs. Mathematically, one cannot analyse non-Markovian processes as if they were Markovian.

Real Ecosystems

We now consider the systems in the lower row of Table 1, starting with a minimal overview of real ecosystems.

Ecological Succession. This is the observed process of change in structure of an ecological community over the medium to long term. For instance after a mass extinction a typical sequence is for a few species of plants and animals to initially return; then successive new organisms arrive, building on what is already there in what Ostrom might want to call multiple nested polycentric layers in analogy to her social systems. In some cases this may be a somewhat predictable succession towards a final ‘climax community’ [3]; but more recent ideas tend to take account of the many historical contingencies involved, including the varied feedbacks through knock-on environmental effects, and see a more unpredictable picture of ‘alternative stable states’ [14]. In the short-term an ecosystem is in a stable steady state, but in the longer term it is somewhat accidental which one of many such possible equilibria it is, and what range of fellow organisms it contains.

Niche Construction. Such theories emphasise that organisms may not be merely accepting or selecting (through moving to) their specific environment; they may also have an active role in changing it [12].

ALife Models of Ecosystem Regulation

Many ecological models incorporate interactions between different species of biota; here we focus on models that explicitly also incorporate environmental variables and their interactions with biota. The context is Gaia Theory.

Gaia Theory. Also known as the Gaia hypothesis, this proposes that mutual interactions between organisms and their natural environmental surroundings (e.g. weather, geology) has resulted on this planet in a complex self-regulating system that tends to maintain conditions conducive to the survival of the organisms [16]. This is controversial, with its apparent appeal to teleology, and some sceptical viewpoints will be discussed below. Daisyworld (DW) models [23, 7] offer a simplified vision of how organisms and environment interact, in some sense cooperatively, in the form of an Artificial Life model. This can be compared to a very basic form of niche construction.

Motivation for Daisyworld Models. These models are not widely known, and where known largely misunderstood [7]. The rationale is to model a number of types of organisms (e.g. one being 'daisies' D) that can survive within a limited range of local environmental conditions (e.g. one such being 'temperature' T). Collective survival of an ecosystem of different organisms means *all* of them are currently viable in their local environment; *robustness* of an ecosystem is measured in terms of how wide a range of perturbing environmental conditions it can survive; e.g. perturbations from an external 'sun' S creating hotter/colder conditions. An organism may have some local environmental effect (e.g. the albedo of a black daisy may raise local temperature), and *complexity* is measured as the number of such different effects within the ecosystem. The key DW result is that more such *complexity* leads to greater ecosystem *robustness*.

We demonstrate this, starting from the simplest ecosystem with a single species; for equations see Appendix A. Figure 2 shows schematically the basic influence of environment T on an organism D. Figure 3 shows the consequence of further adding an effect from the organism D onto the environmental variable T. The consequence is to extend, i.e. widen the range of solar forcing (perturbing external effect S) for which the organism is feasible [7]. Here the effect is positive (e.g. the albedo of a black daisy increases local temperature) the solar feasibility range is extended towards lower values than otherwise. A negative effect (e.g. white daisies tend to reflect heat and decrease temperature) would extend the solar feasibility range towards higher values.

Viability and Feasibility. Discussions in Gaia Theory and Daisyworld modelling have been plagued by a failure to distinguish between these two different but inter-related concepts. Here the term *viability* is used to refer to the deterministic relationship between organisms (e.g. daisies D) and local environmental conditions (e.g. temperature T): when the daisy population as a function of local temperature (as specified by the equations) is positive, this defines a *viability-range*, e.g. [T_{lo}, T_{hi}] defined in degrees Centigrade. By contrast the term *feasibility* here refers to the relationship between organism-viability and external perturbations (here solar forcing S): analysis of the system dynamics should allow us to determine a *feasibility-range* of external perturbations, e.g. solar forcing [S_{lo}, S_{hi}] defined in Watts/m^2, within which a viable (>0) steady-state daisy population is *possible*.

Though S and T are fundamentally different quantities measured in different units, there is a convenient shortcut that allows them both to be measured in T-units of temperature: one may scale measures of solar output S to the temperature T that would prevail *if* all biota was absent, a dead planet. This shortcut is usually taken in the literature, as it is in this paper (e.g. see caption to Figure 2); but it comes at the often-unrecognised conceptual cost of facilitating confusion between *viability* and *feasibility*. Much misunderstanding of Gaia Theory arises from this conceptual confusion, and indeed the present author has been guilty of this himself — hence the need to edit some mentions of 'viability range' from a previous paper [7] to 'feasibility range' in this version. A crucial distinction is that the mapping T→D is single-valued, the *viability* of biota (e.g. daisies) is a deterministic function of local environmental conditions (e.g. temperature); whereas the mapping S→D is potentially more than single-valued, specifically in the regions of a hysteresis loop.

In such regions, for any value of S (or, via the re-scaling shortcut, the corresponding T) there may be either a non-viable D population of D=0, or a viable D population >0. Whether one observes non-viability or viability in such circumstances depends on the contingent history; but the crucial definition of *feasibility*, on which the Gaia Regulation Theorem is based, requires only that at least one such contingent history gives rise to a viable non-zero population at steady-state.

Plus and Minus, Rein Control. Further, if both black and white daisies are potentially available with both positive *and* negative effects on the local environmental variable, temperature, they will collectively expand their joint feasibility range, as seen in Figure 4. This phenomenon depends on some basic assumptions spelt out in Appendix A; each variant, black or white, largely determines its own local temperature but with some 'leakage' between them in their shared environment. In this model, interactions between different 'species' such as D_B and D_W are only mediated via environmental variables, rather than through e.g. direct predation of one on the other. The results here, and developed further in [7], demonstrate that any changes in feasibility range (for D_B&D_W, or D_B or D_W individually) always increase that range and never decrease it.

The expanded feasibility range takes the form of hysteresis loops as in Figure 4b. If the external perturbing force, here S, changes slowly, then which of the upper (viable) or lower (non-viable) arms of such loops is followed depends on which direction they are approached. In this sense history matters.

This is an example of 'rein control' [4, 5]. Clynes [4] observed a pattern when natural organisms exhibit homeostasis in response to external forces threatening viability both from above and below (e.g. both 'too hot' and 'too cold'). Rather than one mechanism responding in two directions, he noted two mechanisms each responding in one direction only — referred to as unidirectional rate sensitivity. Since reins of a horse have this same property, each pulls but does not push, he called this 'rein control'.

This is further related to Le Chatelier's principle [13] as known to chemists and economists. This principle asserts that when any system in equilibrium is disturbed the system will adjust itself so as to (at least partially) nullify the effect of the change. Though posited some 80 years before ideas of rein control were introduced, it is aimed at precisely those forms of balanced dynamic equilibria such as we see in Gaia Theory and DW phenomena that may naturally be described in rein control terms. A practical application of this principle is the use of a *buffer solution* which resists changes in pH when acid or alkali is added. Such buffers may be artificial, i.e. designed by chemists [20], or seen naturally where the *bicarbonate buffering system* regulates pH of blood in humans or other animals [12].

Multidimensional Daisyworld

So far we only considered one environmental variable at a time: say temperature in DW, pH in the buffering examples. What if two or more such variables are simultaneously relevant, e.g. both temperature and pH?

We can answer this within any very simple, abstract class of ecosystem models where (any number of) 'organisms' are modelled by 'hat-shaped' viability functions of (any number of) environmental variables; and in turn the organisms have any effect of any kind, positive or negative, on each or all of the environmental variables. In such cases it has been shown in the 'Gaian Regulation Theorem' [7] that hysteresis loops or buffer zones as illustrated above exist regardless of the dimensionality of any such system. Perturbations in any number of dimensions will tend to be countered so as to widen — and under the stated constraints never lessen — the viability range of any disparate group of organisms in an ecosystem, or of individuals in a corresponding social system.

Gaian Regulation Theorem

In Appendix B we give the most recent version of the proof of this theorem, GRT. A main reason for Gaia Theory being controversial is that there are so many different versions of its supposed claims. Kirchner [10] catalogued and critiqued a number of different spins on Gaia Theory, but none of those correspond exactly to what is presented here as a mathematical theorem: *provided* a dynamical system meets the conditions specified, then *provably* the feasibility range of external perturbations for which a steady state with a viable (>0) population is possible can only be increased, never decreased, by any effects of 'biota' on 'environment'. As a mathematical theorem, the only distinction between 'biota-like' variables B and 'environment-like' variables E is the different form of the equations relating one to the other.

We do not here address the separate scientific question of whether these equations can indeed (plausibly enough) model the actual interactions between biota and nature in the real world. But we note that both the equations of Watson and Lovelock [23], and their demonstrated core results, do indeed correspond to the requirements of the generalised GRT as presented here and its conclusions. GRT fully endorses the core point made in the original DW papers. Many of the subsequent misunderstandings of DW (e.g. see many listed in [10]) relate to the *mathematics* of feedback systems, not to the *science* itself.

GRT summarised

Here we more informally summarise the salient points of Appendix B. We generalise the original DW model to an arbitrary number m of B-variables (usually interpreted as different biota such as black and white daisies, D above); an arbitrary number n of E variables (environmental, such as temperature T above) associated with each B; and the same number n of perturbing forces P (external global perturbations such as solar insolation S above with its effects on temperatures). The interpretations are irrelevant to the mathematics; but the different classes of variables and parameters determine the forms of the simple differential equations defining a dynamical system.

Each B-variable changes over time according to a 'viability-function' of its n local E-variables, that carves up E-space into regions where B tends towards zero (i.e. is non-viable) and the remainder where B tends towards some positive value (is viable). The equations guarantee that B-variables must always remain non-negative. We can summarise this (omitting here any time parameter) as:

$dB/dt = V(E) - B$

Each E-variable that is associated with a B-variable will change over time according to a formula with three salient influences. Firstly, there is an external biasing term that influences E towards some externally fixed global P perturbation. Secondly, there is a 'diffusion' or 'leakage' term that tends to diffuse the E-values of each B towards some average value $\underline{E}$ based on E-values of some or all of the other Bs. Thirdly and crucially, there is an 'effect' term F(), such that the B-variable may have influence so as to either increase or decrease its own local E-values. F() is a continuous, potentially non-linear, function of B subject only to the constraint that F(0)=0, i.e. zero B means zero local effect. We can summarise this as:

$dE/dt = P + \alpha(\underline{E} - E) + \beta F(B) - E$

where α and β lie within the range [0,1].

P are parameters in the sense that — if they do indeed vary — they vary on so much slower a time scale than B and E that the latter have time to reach a stable equilibrium for any given P. The form of the equations guarantees at least one, and potentially more, stable equilibrium/a. The GRT focuses on the conditions — specifically what we define as the *feasible* range of P-values — under which there exist such equilibria with positive B-values, i.e. with viable B.

The claim of Gaia theory is that the effects that biota have on their environment will — under some plausible range of conditions yet to be fully specified — somehow contribute to extending the capacity of the environment to sustain life. This translates within this model to the claim that the effects B-variables have on E-variables acts so as to potentially increase (and never decrease) the *feasible* range of P values, i.e. external perturbations. This corresponds to the core claim of the original DW models [23].

The proof in Appendix B proceeds by comparing the feasible range $\mathscr{F}$ of P values in the standard case (where B affects E) with the feasible range $\mathscr{F}_0$ in the 'null' case where B has no effect on the environment. If $\mathscr{F}$ is bigger than $\mathscr{F}_0$, we may claim that the extended range is attributable to effects B has on E, and call this an example of Gaian regulation. However in practice we adopt a much more stringent condition than $\mathscr{F}$ bigger than $\mathscr{F}_0$, namely $\mathscr{F} \supseteq \mathscr{F}_0$. We cannot prove the more stringent result under all versions of the set of equations sketched out, because there are indeed counter-examples (and we shall show one). But for two simple versions we can indeed prove that the specific constraints entail $\mathscr{F} \supseteq \mathscr{F}_{0.}$ The form these proofs take indicates where we may find our counter-example.

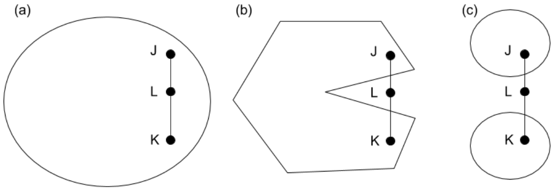

Figure 5. (a) a convex zone ensures that if points J and K lie within, then any intermediate point L also lies within. This is no longer guaranteed for (b) a concave zone, nor (c) for separated zones.

One constraint for part of the proof to work is that we require viability zones to be convex. In other words, any point in E-space that is intermediate between 2 (or more) points lying within such a zone will necessarily also lie within the same zone (Fig. 5). In each case we can motivate the proof that the equation for dE/dt can be recast in the form

$$dE/dt = P + X - E$$

where X (= $\alpha(\underline{E}$ -E) + βF(B)) can be thought of as some potentially moving target, and E moves towards some intermediate value between X and P. If P is fixed at some value within the viability zone, and E starts off at the same value, what constraints on X will ensure that E never leaves the viability zone?

The first rather trivial Scenario I we can prove is where diffusion effects are zero (α =0). This means that there are zero interactions between the local environments of any one bio-variable and any other; in effect each bio-variable exists in its own isolated world, and in that case we can quickly show (Appendix B) that $\mathscr{F} \supseteq \mathscr{F}_0$. The proof can be summarised as taking

$$dE/dt = P + \beta\, F(B) - E \qquad \text{(since } \alpha = 0)$$

and noting that this indicates a vector of change of E towards P + β F(B). If P is within B’s viability zone, and one initialises E at the same value P, then subsequent dynamics will move E but can never take it across the zone boundary — because at that boundary B=0 and hence F(B)=0. So wherever E finishes up, it must be within B’s viability zone.

The second Scenario II we can prove is where we allow diffusion effects (so that the different Bs interact) but require all their different viability zones to cover the identical convex region of E-space. This means that we reintroduce the term $\alpha(\underline{E}$ -E); but since $\underline{E}$ is derived from averaging across different Bs, then so long as the Bs are viable then the convexity entails that $\underline{E}$ will also lie within the same shared viability zone. Once again this means that the dynamics cannot take E across the zone boundary.

The details of the proof in Appendix B confirm that it is a sufficient — but not necessary — requirement for all the Bs to share the same convex viability zone; this guarantees $\mathscr{F} \supseteq \mathscr{F}_0$. We note that this points to where the proof may break down: when viability zones are not identical, and the diffusion effect may be strong enough to ‘pull’ an E-value outside a viability zone. This motivates the minimal counter-example shown in Fig. 6, with a single E-variable and two B-variables. Using the numbers shown, we can see that — although individually in the absence of diffusion, B_1 and B_2 do extend their P-feasibility ranges — collectively in the presence of diffusion B_1&B_2 do not.

Perhaps surprisingly, it is somewhat difficult to construct such a counter-example. Fig. 6 demonstrates an ‘exception that proves the rule’. The conditions on convex shared viability zones that do guarantee Gaian regulation are sufficient but not necessary for the proof to hold, and it turns out that in practice many of these constraints can be relaxed whilst still providing such regulation. Analysis [5, 7], together with observation and common sense, support heuristics that large B-effects on E (whether positive or negative) are more likely to extend feasibility range than small effects; and where there are many B variables interacting, greater extension of feasibility range is promoted by having *diffusion* or ‘leakage’ terms at intermediate values. Too little diffusion results in B variables having little interaction, whereas too much diffusion means their behaviours tend to merge towards one global average where positive and negative effects cancel each other out. This observation motivates the use of intermediate diffusion or leakage values in both the simple 2-B version of the original DW (see Appendix A) and in the following illustrative multi-B example (elaborated in Appendix C).

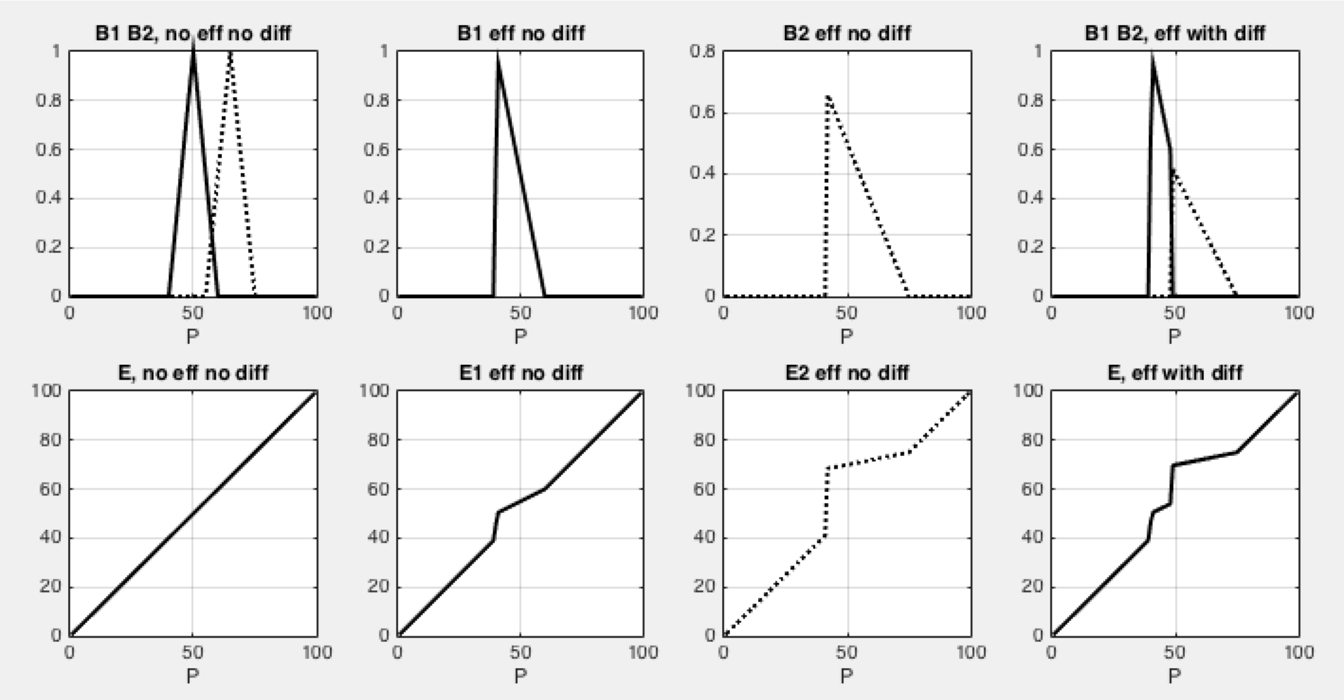

Figure 6. A counter-example where the addition of interacting effects *does indeed* reduce the feasibility range of B_1&B_2, with a single environmental variable E. B_1 and B_2 have witch's hat viability zones of radius 10, centres 50 and 65. Hence in absence of effect or diffusion (column 1), feasibility ranges are [40,60] and [55,75], with an overlap [55,60]. B_1 effect size is +10, meaning (column 2) its self-effect does not extend viability range [40,60]. B_2 effect size is +40, meaning (column 3) its range extends to cover [42,74] with its self-effect. However if (column 4) these effects are shared with a diffusion rate of 0.5, the interactions means B_1 and B_2 mutually exclude each other, they now have zero common range of feasible P values.

Illustration of Multidimensional Gaian Regulation

As an illustrative example of Gaian regulation in some abstract multidimensional domain, Figure 7a shows 8 groups of 8 species in clusters of narrow preferences for 3 environmental variables. Within each group, there is indeed the shared convex viability zone discussed in the proof. In the absence of DW feedbacks at most one such group could be viable since the small group-viability spheres do not intersect (only P and V spheres shown here). If we add DW effects, different for all 8 members within each group, then when an external perturbation happens to pass the neighbourhood the whole group of 8 becomes jointly viable with a feasibility radius greatly expanded (from 0.05 to 0.218 for effect size 0.4; details in Appendix C). Given the technical definitions of *feasibility* and *viability* given above, we may call the expanded region a *feasibility region* (based on the axes being *perturbation* dimensions); whilst the original unexpanded spheres are *viability regions* (based on *environmental* variables providing the dimensions). The expanded feasibility spheres (e.g. V-sphere in Figure 5b shows the expansion for an effect size of 0.4) may now overlap and (depending on environmental history) several such groups may become simultaneously viable; feasibility regions may overlap even where the viability regions do not — indeed this lies at the core of the DW phenomenon. If the effect size were increased (from the 0.4 shown in Fig. 7b) to 0.8, the mid-value perturbation C (0.5,0.5,0.5) would be within all 8 such potentially expanded feasibility spheres, thus allowing all 64 (8x8) species with diverse environmental limits and diverse effects to be simultaneously viable — provided that contingent history did accumulate all these buffering steps.

Hysteresis and History

This illustrative proof of principle still only has 3 dimensions of environmental/perturbation variables, and is symmetrically set up to demonstrate an effect. The GRT is as valid for 103 dimensions as it is for 3, and the motivation for using 3 here was only that it is easier to visualise in a diagram. Real systems will typically have more dimensions and be highly asymmetrical and locally varied, with convoluted overlaps of basins of attraction. Nevertheless we can see that different perturbation trajectories may result in very different ecosystems. Trajectories matter, history matters — and the choice points relate directly to hysteresis effects. Hysteresis and history may be etymologically largely unrelated, but conceptually there is a strong connection.

Such meandering paths through ecosystem-space can be compared with meandering evolutionary paths through DNA-space, and some degree of resemblance is not entirely accidental. From a high-level perspective the viability functions of DW can be related to the survival focus of Darwinian evolution. The natural settlement into attractors of the broad class of dynamical systems that is multidimensional DW relates directly to the natural selection of Darwinian evolution. Indeed the latter may be seen as a special case

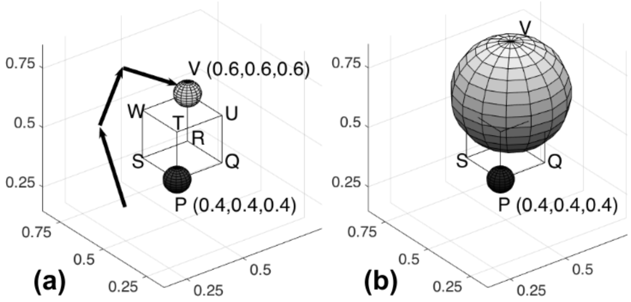

Figure 7. (a) 3 dimensions of external environmental perturbations. P is group of 8 species, preferred env. (0.4,0.4.0.4), viable within radius 0.05 of this as shown by P sphere. Similar groups centred on Q, R…W, at corners of cube. Arrows show a possible trajectory of external perturbation. (b) This passes through viability zone of group V, DW effect consequently expands its viability radius to 0.218. See Appendix C.

of the former. They both have surprising and counter-intuitive consequences; for instance an increase in an effect that *increases* the range of DW robustness usually *decreases* a conventional measure of Darwinian fitness [7]; feasibility and viability are subtly different.

Where an A-Historical Analysis Differs

The analysis of ecosystems in terms of DW, with robustness associated with complexity, runs counter to the expectations of some. One influential analysis [17] of an even broader class of 'any large complex system' (that includes multidimensional DW) purports to contradict it, proposing that, after some critical number of variables is exceeded, such systems are inherently unstable. Three mathematical flaws in this analysis have been previously exposed [6]. We here go further in identifying these flaws as arising from an a-historical analysis that resembles the a-historical analysis of IPD [19, 22].

May [17] picks out an arbitrary equilibrium point of a large complex system and analyses its properties. This arbitrary choice, together with other explicit or implicit assumptions he makes, allows one to draw general conclusions; as the system gets larger, the chance that this specific equilibrium is stable tends towards the vanishingly small. This part of May's argument resembles creationists' arguments about the improbability of the 'irreducibly complex'. But even though the probability of an arbitrary lottery ticket being a winner may become arbitrarily small as the lottery itself gets arbitrarily big, this does not in itself prevent there being a winning ticket, or indeed many such.

A dynamical system left to its own devices will naturally head towards a stable equilibrium, such a 'winning ticket'; GRT assumptions ensure there is indeed at least one stable one, and any unstable equilibrium will only be briefly observed. As external conditions change, such a system inevitably passes through a sequence of metastable states. Thus any observed equilibrium is almost inevitably a stable one; which equilibrium it is depends on the history of the system. May's analysis [17] of a generic a-historical equilibrium state has little relevance for the analysis of specific, observed, historically contingent equilibria [6]. Likewise the analysis by Press and Dyson [19] of extortionate ZD strategies for IPD, or by Stewart and Plotkin [22] of generous ZD strategies, has little relevance for historical contingent strategies such as TFT_C or TFT_D.

Conclusions

Crudely speaking, biology equals physics (and chemistry) plus history — stability in the short term plus the contingent context arising from an extended history of (meta-)stability. More elegantly put, "Biology has always occupied a middle ground between the determinism of classical physics and the uncertainties of history" [21]. When the physics of short-term stability is the focus of attention to the exclusion of contingent history, key concerns that can characterise complex systems can be missed.

It may be more than a coincidence that Press, Dyson, Stewart, Plotkin and May, variously cited and criticised above, all come from physics backgrounds. Another physicist, Rutherford, is quoted [2] as saying "All science is either physics or stamp-collecting". If the latter is interpreted as contingency, it need not be taken as derogatory; this is not only important for understanding real biology and social science but equally so for Artificial Life models of these.

In biological systems internal DNA is one obvious marker of a history, but other external markers may also be crucial. In polycentric social contracts [18] there may be multiple overlapping simultaneous systems of governance; likewise in polycentric organisms, polycentric ecosystems. Adaptations (and neutral changes) in any one system layer are within (and constrained by) the contingent current context of the others. Complexity of the whole arises through such adaptive/neutral trajectories through history, and cannot be explained a-historically.

A specific novel observation in this paper, apparently not noted by other commentators, is that the recently discovered extortionate ZD strategies in IPD [19], together with their generous cousin strategies [22], have very little relevance to any biological or social studies of cooperation because they are all avowedly a-historical. Their Markovian assumptions are mathematically powerful but implausible as models of reality. The same applies to May's analysis of large complex systems [17].

The Gaian Regulation Theorem has been presented here in its most general form to date. As a mathematical analysis of a very broad class of dynamical systems, it may be applicable far beyond its origins in Daisyworld models. The different trajectories available at the hysteresis loops central to such systems are key to their capacity for accumulating history, and the GRT points to relationships between adaptive stability and history.

Successful real social systems and ecosystems have a history of adapting to circumstances, and this gives context to their current stability. Artificial Life models should reflect this, and there are currently many promising research areas that give scope for developing what are often currently deficient a-historical models too as to take account of such contingency. History matters.

Appendix A

Figure 4 shows 'black' and 'white' daisies, D_B and D_W, and respective local temperatures T_B, T_W [5]. Figure 3, using D_B only, is similar except that D_W is clamped to 0.

Daisy viability w.r.t. local temperature is based on a 'hat-shaped' function H(T) with (Figure 2) peak value 1.0 at T_{opt} reducing to zero outside some limited viability range. Results are not qualitatively changed by different hat shapes.

(A1) $H(T) = \max(0, 1 - \mathrm{abs}(T_{opt} - \alpha T))$

Parameter α sets slope of hat. hence radius (=1/α) of daisy-viability in terms of its local temperature. Parameter β sets the rate at which daisy-viability moves towards the hat-function:

(A2) $dD_B/dt = \beta\ (H(T_B) - D_B)$

(A3) $dD_W/dt = \beta\ (H(T_W) - D_W)$

The local temperature T_B, of black daisies D_B is based on the solar insolation S, altered (i) by *positive* influence from the black daisies, and (ii) by equilibration towards T_W. T_W is conversely affected, white daisies have *negative* effect. On the assumption that temperatures settle faster than rate of change of Daisies we can use the steady-state values as in [5]. Using T′ for intermediate values of T, phase (i) is:

(A4) $T'_B = S + \gamma\ D_B$

(A5) $T'_W = S - \gamma\ D_W$

where γ parameterises the effect size for black/white daisies increasing/decreasing their own local temperatures. Phase (ii) gives the final temperature T as a compromise between each individual T′ and their average current values; there is some 'leakage' [5], here parameterised via δ (for $0 \leq \delta \leq 1$), between temperatures of black and white daisies:

(A6) $T_B = \delta\ T'_B + (1- \delta)(\ T'_B + T'_W)$

(A7) $T_W = \delta\ T'_W + (1- \delta)(\ T'_B + T'_W)$

If we choose δ=0.5, then algebraic manipulation shows that equations (A4,A5) together with (A6,A7) can be replaced by:

(A8) $T_B = S + \varepsilon\ (3\ D_B - D_W)$

(A9) $T_W = S + \varepsilon\ (D_B - 3\ D_W)$

where for convenience we substitute ε (= γ/4) for parameter γ.

Equations (A1), (A2,A3) and (A8,A9) can be simulated computationally by choosing some specific value for S, and running these equations from starting values for D, T, until steady-state is reached. In hysteresis regions, the end-states reached will depend on the starting states. To plot one branch of each hysteresis loop, S should be initialised at a low value, and the computation run until D, T reach steady-state. Then S is incremented slightly, keeping *current* values of D, T as new starting values for the next run; this is further repeated, through to high values of S. If the process is then reversed, moving from high S to low S, the other branches of the hysteresis loops can be plotted. In Figure 4b, the feasibility of D_{B+W} is plotted as: IF (D_B>0 AND D_W>0) plot 1, ELSE plot 0.

Appendix B

We here develop the formal analysis of the Gaian Regulation Theorem of [7] to be even more general, covering arbitrary numbers of variables. The proof requires certain constraints that we show are sufficient (but not necessary).

We use 3 classes of variable, B, E and P, that correspond to Biota, Environmental variables and external Perturbations, generalised to m biota B_i (i=1..m), and n types of environmental variables (env-var) and perturbations. This gives us P_j (j=1..n); and since each biota has its own local env-var, we have E_{ij} (i=1..m, j=1..n).

We assume that perturbations are slow-moving enough to act as fixed parameters for long enough for the other variables to reach equilibrium, as governed by the equations:

(B1) $\tau_i\ dB_i/dt = V_i(E_{i,j=1..n}) - B_i$
(m eqns for i=1..m)

(B2) $\tau'_{ij}\ dE_{ij}/dt = M_{ij}(P_j)\ + \alpha D_{ij} + \beta\ F_j(B_i) - E_{ij}$
(mn eqns for i=1..m, j=1..n)

where M(), D() and F() are respectively Monotonic, Diffusion and Effect functions to be defined below; α and β (in range [0,1]) parameterise degrees of diffusion and effect; τ and τ' are positive parameters setting timescales for the trajectories of B_i and E_j towards their respective nullclines (given by dB/dt=0, dE/dt=0):

(B3) $B_i = V_i(E_{i,j=1..n})$
(B4) $E_{ij} = M_{ij}(P_j) + \alpha\, D_{ij} + \beta\, F_j(B_i)$

$V_i()$ is any continuous non-negative bounded *Viability* function of any or all the E; $V_i()$ divides E-space into a region $\mathcal{V}^+{}_i$ where $V_i()>0$ (B_i is viable) and the remainder $\mathcal{V}^0{}_i$ where $V_i()=0$ (B_i is non-viable). For the purposes of this proof we are going to add the quite strong constraint that each such viability region $\mathcal{V}^+$ is convex, as illustrated in Figure 5. $M_{ij}()$ is any continuous monotonic (for example linear) function of external perturbations that acts as an *external biasing* term. D_{ij} represents a diffusion term that tends to equilibrate E-values towards $\underline{E}_i$ some average of different 'neighbouring' environmental values; for any given i, $\underline{E}_i = \sum w_{ij}E_{ij}/\sum w_{ij}$ (summing over j) where all weights are in range $0\leq w<1$, and $D_i = \underline{E}_i - E_i$. For example, in the original DW these terms mediate the diffusion of heat between populations of black and white daisies. $F_j()$ is any continuous bounded *Effect* function of any or all the B, constrained only by the requirement that $F_j(0)=0$ (zero B means zero effect).

The simplicity of this family of ODEs means that there are only point equilibria, where the nullclines intersect; each equilibrium is either stable or unstable with only the stable ones acting as attractors. Since the nullclines are continuous, it can readily be shown that neighbouring equilibria along any intersection of nullclines must alternate between stable and unstable. Because the nullcline functions are bounded, there must be at least one stable point attractor; and in fact there will be some odd number 2a+1 (for some $a\geq 0$) of equilibria of which a+1 are stable, the rest unstable [7].

We are interested specifically in any stable equilibria where biota are viable, B>0; we may be considering a single B, several, or all of them collectively. We define as *feasible* those regions $\mathcal{F}$ of P-space where there exists at least one stable equilibrium with B viable. We do this under two different scenarios: firstly the standard scenario where $\mathcal{F}$ defines the feasible parts of P-space with the B-effects as specified above, and secondly the null scenario where $\mathcal{F}_0$ defines the feasible parts of P-space when the B-effects are all nullified or turned off. We can then look at the difference between $\mathcal{F}$ and $\mathcal{F}_0$ as defining the difference in feasible P-space attributable to those B-effects being non-zero. We now proceed to prove for two specific scenarios that $\mathcal{F} \supseteq \mathcal{F}_0$; in other words the B-effects can only expand the range of P-space feasibility (as compared to the range with null B-effects) and never decrease that range.

Scenario I, no diffusion. If we set α=0, there are no diffusion terms and indeed no interaction at all between the different Bs. Hence in this rather trivial scenario we can treat each one independently. For any one B, eqns (B2, B4) become (omitting i):

(B5) $\tau'_j\ dE_j/dt=M_j(P_j) + \beta\, F_j(B) - E_j$

and at equilibrium (B6) $E_j = M_j(P_j) + \beta\, F_j(B)$

Within a j+1 dimensional phase space, comprising the j dimensions of E (or P) plus the one dimension of a single independent B, the B-nullcline of eqn (B3) comprises a manifold whose B>0 values define a viability-zone $\mathcal{V}^+$ for B. The E-nullcline of eqn (B6) is a line passing through E=M(P) when B=0. Joint equilibria for both B and E occur at point(s) where the line intersects the manifold. Consider first the null case where β=0, and any external perturbation P* for which there is a viable (B>0) stable equilibrium. It follows from (B6) that that equilibrium value $E_0{}^*=M(P^*)$, and $E_0{}^*$ lies within the viability zone $\mathcal{V}^+$. Moving on now to the standard case where β=1, with the same external perturbation P*, we see that the E-nullcline of (B6) becomes E = M(P*) + F(B). Let the first point at which this line intersects the B-nullcline of (B3) be at $E=E_+{}^*$. We have

(B7) $E_+{}^* = E_0{}^* + F(B)$

But $E = E_0{}^* + F(B)$ defines a continuous line, passing through the point ($E=E_0{}^*$, B=0) and necessarily intersecting the B-nullcline manifold where B>0. So $E_+{}^*$ must also lie within the viability zone $\mathcal{V}^+$.

In other words, any P* that provides a viable (B>0) stable equilibrium in the null case will necessarily also provides a viable stable equilibrium in the standard case: for this no-diffusion Scenario I, for each B independently (and hence for all collectively) $\mathcal{F} \supseteq \mathcal{F}_0$.

Scenario II, with diffusion, but with a common viability zone. Here we wish to consider positive α, e.g. α=1, when in consequence the diffusion of environmental variables means that the Bs can no no longer be treated independently. We can still prove the desired result, provided we add the significant constraint that all the different Bs share the same common convex viability zone $\mathcal{V}^+$.

The proof is by *reductio ad absurdum*, considering what would happen if we choose to vary α continuously from 0 to 1. We know that at α=0 we have the desired result since this is Scenario I. Suppose at some intermediate value of α, as it is increased, the result first fails to hold true: one at least of the Bs fails to have a viable stable equilibrium in the standard case. At that moment, for each B, their E-values must be either within $\mathscr{V}^+$ or (for Bs on the point of becoming non-viable) at the very boundary of $\mathscr{V}^+$. Hence at that moment any weighted average $\underline{E}$ of some or all the Bs must — because of the convexity of $\mathscr{V}^+$ — itself lie within $\mathscr{V}^+$. In consequence we can go through the same steps of the proof as for Scenario I, considering one specific B on the point of becoming non-viable, except that at equation (B7) our E-nullcline now includes the diffusion term:

(B8) $E = E_0^* + D + F(B)$

Because D is defined in terms of $\underline{E}$, $(E_0^* + D)$ must be within $\mathscr{V}^+$. So the same arguments as in Scenario I will hold, replacing E_0^* with $(E_0^* + D)$, and again the E-nullcline intersects the B-nullcline manifold where B>0. So again E_+^* must also lie within the viability zone $\mathscr{V}^+$. But this contradicts our supposition that we are considering a value of α where this ceases to be true.

Thus in Scenario II, using our constraint a common convex viability zone $\mathscr{V}^+$ shared between all the Bs, the diffusion terms will not shift any equilibria out of such zones; again we have $\mathscr{F} \supseteq \mathscr{F}_0$.

We should note from the details of the proof that it requires the null B-effects not merely to refer to any B-effect that *directly* affects the viability of itself, but also to B-effects that may only do so indirectly via other Bs.

Appendix C

Figure 7 shows 3 dimensions of external env. perturbations. Viability of group of 8 species at P is 1.0 at (0.4,0.4,0.4), decreasing linearly to 0.0 at radius (Euclidean distance) 0.05. Each species has different +/- effects on 3 respective env. variables (2^3 = 8 variants); signs differ, but effect size is always 0.4. The other 7 groups (Q,…,W) are formed similarly.

Effects of a P-species are multiplied by their viability and have two local contributions: half serves to shift the P-group local env. away from the perturbing force (and is thus shared with other P-members; 'leakage' or 'diffusion'); and half shifts the species-specific env. away from the P-local env. This choice is motivated by the observation that extended feasibility range is promoted by such intermediate diffusion values. Over a trajectory of env. perturbations, at each point 20,000 computational iterations altered viability by 0.001 and local env. variables by 0.005 of their indicated shift. This smoothing of dynamics, together with the inheritance of previous env. values as perturbations changed, avoided numerical instabilities. A species was considered extinct if viability<0.01.

An effect size 0.4 expanded feasibility radius of each group from 0.05 (the basic viability radius with null effects) to 0.218; an effect size 0.8 is seen to expanded feasibility radius further to 0.35.

References

[1] Axelrod, R., (1984). The evolution of cooperation. Basic Books, NY.

[2] Birks, J. B. (1962). Rutherford at Manchester. Heywood, London.

[3] Clements, F. E., (1916). Plant succession; an analysis of the development of vegetation. Carnegie Institute of Washington.

[4] Clynes, M., (1969). Cybernetic implications of rein control in perceptual and conceptual organisation. *Annals of the New York Academy of Sciences* 156:629-670.

[5] Harvey, I., (2004). Homeostasis and rein control: from Daisyworld to active perception. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T. and Watson, R. A. (Eds.), *Proceedings 9th International Conference on Simulation and Synthesis of Living Systems, ALIFE 9.* pp 309-314, MIT Press, Cambridge, MA.

[6] Harvey, I., (2011). Opening stable doors: complexity and stability in nonlinear systems. In Lenearts, T. et al., (Eds.), *Advances in Artificial Life, ECAL 2011,* pp 805-812, MIT Press.

[7] Harvey, I., (2015). The circular logic of Gaia: fragility and fallacies, regulation and proofs. In Andrews, P., Caves, D., Doursat, R., Hickinbotham, S., Polack, F., Stepney, S., Taylor, T. and Timmis, J. (Eds.), *Proceedings European Conference on Artificial Life 2015*, pp 90-97, MIT Press.

[8] Harvey, I., (2016). Social systems and ecosystems: History matters. In Gershenson, C., Froese, T., Siqueiros, J.M., Aguilar, W., Izquierdo, E. J. and Sayama, H., (Eds.), *Proceedings of the Artificial Life Conference 2016,* pp 418-415, MIT Press.

[9] Hobbes, T., (1651). Leviathan. Andrew Crooks (publisher), at the Green Dragon in St. Pauls Church-yard, London.

[10] Kirchner, J. W., (1989). The Gaia hypothesis: can it be tested? *Reviews of Geophysics,* 27(2):223-235.

[11] Krieg, B. J., Taghavi, S. M., Amidon, G. L., Amidon, G. E., (2014). *In vivo* predictive dissolution: transport analysis of the CO2, Bicarbonate *in vivo* buffer system. *Journal of Pharmaceutical Sciences,* 103(11):3473-3490.

[12] Laland, K. N. and Sterelny, K., (2006). Perspective: seven reasons (not) to neglect niche construction. *Evolution,* 60(9), 1751-1762.

[13] Le Chatelier, H., (1884). Sur un enoncé général des lois des équilibres chimiques. *Compts Rendu*, vol. 99, pp. 786-789.

[14] Lewontin, R. C., (1969). The meaning of stability. In Woodwell, G.M. and Smith, H. H. (Eds.) *Brookhaven Symposia in Biology,* 22:13-23, Brookhaven NY.

[15] Lindgren, K., (1991). Evolutionary phenomena in simple dynamics. In Farmer, J. D., Rasmussen, S. and Taylor, C., (Eds.), *Artificial Life II,* pp 295-312. Edison-Wesley, Redwood City, CA.

[16] Lovelock, J. E. and Margulis, L., (1974). Atmospheric homeostasis by and for the biosphere; the Gaia hypothesis. *Tellus,* Series A 26(1-2):2-10.

[17] May, R. M., (1972). Will a large complex system be stable? *Nature* 238, 413-415.

[18] Ostrom, E., (1990). Governing the Commons: the evolution of institutions for collective action. Cambridge University Press.

[19] Press, W, H. and Dyson, F. J. (2012). Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences.* 109(26), 10409-10413.

[20] Scorpio, R. (2000). Fundamentals of acids, bases, buffers and their application to biochemical systems. Kendall Hunt, Dubuque, IA.

[21] Smith, T.F. and Morowitz, H. J., (1982). Between history and physics. *Journal of Molecular Evolution,* 18(4), 265-282.

[22] Stewart, A. J. and Plotkin, J. B., (2013). From extortion to generosity, evolution in the Iterated Prisoner's Dilemma. *Proceedings of the National Academy of Sciences,* 110(38), 15348-15353.

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