2019 In Press. To appear in *Adaptive Behavior* special issue celebrating John Stewart

Neurath's boat and the Sally-Anne test: Life, Cognition, Matter and Stuff

Inman Harvey
Evolutionary and Adaptive Systems Group, University of Sussex
inmanh@gmail.com

Abstract

Making sense of the world around us is likened to the task of staying afloat on a stormy sea whilst rebuilding our craft of ideas and concepts as we go. This metaphor is pursued through successive stages of cognitive development, and more sophisticated appreciation of multiple perspectives; from pre-theoretical to folk science to the theoretical, from individual to social to inter-subjective agreement. This inescapably generates reflections on the relationships between embodied and situated Life and Cognition.

Keywords

Cognition, life, autopoiesis, epistemology, theory-theory, representation, situatedness.

Introduction

I first encountered John Stewart in December 1991, when he was presenting a paper "Cognition = Life" at the First European Conference on Artificial Life (ECAL1991) in Paris. Stewart had had collaborations with both of the organisers of ECAL1991, Varela and Bourgine (1992), and his was one of several contributions that signaled the distinctive change of emphasis that ECAL brought to Artificial Life. The original Alife conferences, with origins in Los Alamos and Sante Fe, tended to project the assumptions of physicists and computer scientists onto projects of synthesising life. In so far as any philosophical assumptions were evident, they tended to mirror the computationalist prejudices of classical AI: if intelligence and cognition were to be taken as some form of computation, maybe life could be treated likewise.

In contrast, ECAL brought both more input from biologists and more explicit attention to philosophical issues, with a distinctive flavour that owed much to Varela. A different perspective was brought to such questions as 'What is Life?', with influences from autopoiesis and constructivist alternatives to objectivism. Here I focus on Stewart's contribution at ECAL1991 (Stewart, 1992) and subsequent related papers (Stewart, 1996; Bourgine and Stewart, 2004; Stewart, 2010). Within a broadly autopoietic framework, he asked what is the relationship between Life and Cognition, is it equality or some form of entailment or what? My answers will be 'no, it is definitely not equality' and 'loosely yes, they do in some sense entail each other'.

In reaching my answers, that may not be too far from those of Stewart, I shall be taking a rather different route. I shall be appealing to the metaphor of Neurath's boat, likening our attempts to make sense of the world to the task of staying afloat on a stormy sea whilst rebuilding our craft of ideas and concepts as we go. Not only is Life embodied and situated, but our reasoning in making sense of the world is also embodied and situated. Stewart (2010) has been concerned to emphasise a continuity of cognition from "low-level" sensorimotor issues to so-called "high-level" cognition; the metaphor of successive versions of Neurath's boat appeals to very similar intuitions.

Much of this paper will be taken up with non-technical sketches, largely based on this metaphor, of how our pre-theoretical ideas of biology and cognition may be built up. Such pre-theoretical sketches are non-rigorous, and open to revision. But the important point to recognise is that any rigorous theoretical analysis must start from somewhere like this, it cannot start from a *tabula rasa*, or a disembodied and unsituated 'view from outside'. The current end-point of the cognitive trajectory shown here is a brief version of my own philosophical position, that has been presented elsewhere rather more thoroughly (Harvey, 1996; 2000). The sketchiness here is because the arc of the trajectory is more of the focus of attention than the end-point.

Within such a work-in-progress, revisable framework, we find that there can be a rather simple and unproblematic relationship between life and cognition, with little conflict between the pre-theoretical and the theoretical: 'cognition' basically covers the relationships between a living organism and what concerns it in its world. So though Life ≠ Cognition, the language of cognition entails a context of such a living organism and its world — or, by metaphorical extension, something that can be mapped onto such a context. We focus on how a living organism does indeed relate to its world, but our analysis does not start by positing axioms and then building a theoretical framework. Instead, we take a historical perspective.

Neurath's boat

This is a philosophical metaphor that emphatically should not be confused with Theseus' boat. The latter is presented as a conundrum: if Theseus' boat has repairs done, over an extended period of time, until eventually every plank, rope and sail has been replaced, is it the same boat as the end as it was at the beginning? This is a disembodied armchair exercise, and the armchair is situated nowhere near the sea; the rebuilding is in a dry dock. By contrast, Neurath's boat (originally from Neurath (1921) but here in a 1944 version) places a similar task in a very different context:

"Imagine sailors, who, far out at sea, transform the shape of their clumsy vessel from a more circular to a more fishlike one. They make use of some drifting timber, besides the timber of the old structure, to modify the skeleton and the hull of their vessel. But they cannot put the ship in dock in order to start from scratch. During their work they stay on the old structure and deal with heavy gales and thundering waves. In transforming their ship they take care that dangerous leakages do not occur. A new ship grows out of the old one, step by step—and while they are still building, the sailors may already be thinking of a new structure, and they will not always agree with one another. The whole business will go on in a way that we cannot even anticipate today. That is our fate." (Neurath, 1944; p.47)

This metaphor was largely influential on, and popularised by, Quine (1950, 1960). He saw this as a picture of how we develop both our everyday language and our more specialised technical and scientific language for the world about us. We cannot start afresh from some disembodied armchair or some dry dock, we have to start from where we are now. Our ideas are always works-in-progress, they have to remain viable as they progress. Viability is measured in terms of empirical testing both against our individual subjective experience and — if we want to meet scientific criteria for objectivity — against inter-subjective challenges, for repeatability in different contexts. We run the risk here of pushing this metaphor too far, but at a minimum we can point to some crude parallels between the life and death of a body of ideas and the life and death of an organism; between the evolution of ideas and the evolution of life-forms.

A trajectory through successive drafts

The following sections are structured in the form of just one possible trajectory through successive versions of such a ship being rebuilt upon the ocean. Primarily this ship is presented as a metaphor for the developing cognitive framework of an individual person, from infancy to adulthood and then to being a cognitive scientist. But there are potential parallels to be drawn with the development over centuries of cognitive science itself. At a much larger scale, without taking Haeckel's slogan of 'ontogeny recapitualises phylogeny' too literally, there are at least comparisons to be made with the presumed evolutionary trajectory that led to *homo sapiens*.

Much of this trajectory will approximate my own cognitive development, but only roughly. At each stage of this journey, we ask what people at that stage might see as the relationship between Life and Cognition.

First and Second Drafts: Pre-theoretical

As a 21st century human, with childhood acculturation in a specific background, and adult exposure to various cultural and intellectual influences, I cannot start from a *tabula rasa*. But I can sketch out a (naive) Just-So Story as to how my initial naive concepts relating to life and cognition were formed. This is not intended to be complete or thorough, but rather to contextualise the saga of how the Neurath ship of cognitive theory developed from early beginnings.

Infant physics, infant biology, infant cognition

From birth, my parental nurturing environment colluded with my instincts and motives to develop my relationships with significant others, and with the basic physics of my world. I learnt the basics of gravity — not *knowing-that* 'it works like this' but *knowing-how* to move my limbs, and to manoeuvre the spoon right-side-up so the food went in my mouth and not on the floor. I went through some fairly predictable developmental stages (Piaget, 1977), some commonly repeated versions of Neurath's boat-rebuilding. Between 1 and 2 years old, I started to be aware that objects may continue to exist even when temporarily not visible. This notion of object permanence formed a major plank in the folk-physics sector of the cognitive boat framework I was developing on my voyage through life.

Later, my language skills started to develop, and my interpersonal relationships broadened beyond my immediate family. People were clearly different from rocks, they responded more directly and actively to me and my concerns, they were more complex. I realised that they were aware of things as I was, and slowly I progressed from an egocentric assumption that their awareness was identical to mine to more sophisticated versions of a 'theory of mind' (Baron-Cohen et al., 1985) that took account of different perspectives.

'Theory of mind' is a misnomer in this context, as from the subject's perspective it is still pre-theoretical. A person may be competent with interpersonal relationships, or with riding a bicycle, without being able to articulate any theory of mind, or theory of bicycle dynamics. Nevertheless, the developmental stage of starting to appreciate that other people have their own perspectives is a significant rebuild of the cognitive craft. Object permanence corresponded to the notion of an objective world independent of us, but this new development shows that unexamined objectivity is not everything; this new competence in relating to people adds subjectivity, in at least this sense of recognising different perspectives on that objective world. Before object-permanence, clouds and rocks fell into either category: 'present and visible' or category: 'absent'. With object-permanence there is a new category: 'present even though not currently visible to me'.

Children interacting with other people start to appreciate how relational cognition is. The Sally-Anne test (Baron-Cohen et al., 1985) involves play-acting with two dolls Sally and Anne, and asking the child where Sally thinks an object such as a sweet is, when Anne (out of sight of Sally) has been shown to hide the sweet in a basket. Infants less than 4 years old typically assume that Sally will have the same knowledge as Anne has (and the infant has). It is only after this age that children start to realise that such knowledge is not objective and universal but is relative-to-a-cogniser. This realisation is often glossed as a 'theory of mind'. But it can also be attributed to squirrels and birds hiding food items out of sight of conspecifics (Steele et al., 2008); hence it does not seem to need the prior development of language.

Folk physics, folk biology, folk cognition

By this stage, moving beyond infancy we have a second draft: the rudiments of folk physics, folk biology and folk cognition. Still not yet, of course, in any theoretical *knowing-that* sense, but

rather still in terms of *knowing-how* to behave appropriately; both in the sensorimotor context of objects and gravity and dynamics and in the interpersonal context of living people (and by extension pets and other animals) with their own cognitive relationships with the world. In terms of the Neurath's boat metaphor, this is not the starting-point of our understanding of the world and our place in it. It is the end-point of a pre-theoretical phase, providing an adequate flotation device, a basic craft for sheltered waters. It can be taken as the starting-point of a theoretical phase, as we develop more sophisticated models and a more scientific approach to physics, biology and cognition. What superstructure can be safely built as we go, and will some of our original planks prove doubtful and need replacing?

Third Draft: using representations, the start of theory

Here we move into the realms of language, of representation and modelling, of *knowing-that* rather than just *knowing-how*. Theorising with the sophisticated use of representations is a peculiarly human cultural practice, that when done properly can produce impressive practical results, and dramatically expand the navigational possibilities for Neurath's boat.

How do representations change everything?

The use of the word 'ball' starts relatively simply. The baby and her mother, exploring and playing with sounds, find that this pattern of sounds 'b-a-l-l' as used by each influences the behaviour of the other, and with experience that influence can be used purposefully. The choice of that particular pattern of sounds was influenced by the mother's linguistic community. As the child develops and its own social community expands, the sophistication of use of the word expands, until 'bring the ball tomorrow to play a game' becomes a meaningful sentence. Just as smoke is indicative of the presence of fire (for physical and chemical reasons), the word 'ball' becomes indicative of a ball (for social and linguistic reasons).

The world itself has expanded: whereas before there were just physical 'things' (such as water and balls) and 'people' (living things, e.g. also extended to animals), now there are 'physical things' and 'people' and 'words'. And words are neither (physical) things nor people, they are relational (abstract) things that require a social context. Words are part of a larger class of representations that includes written as well as spoken language, drawings and even dance. The use of representations is essential for all aspects of modern human society, including of course theorising, and science and philosophy. If a rock on its own involves no relational terms, a rock plus different observers introduces observer-dependent relationships; and with representation-using participants we maybe now have 'relationships-squared', relations about relations.

Animals can use sound signals for identification purposes or for warnings, in ways that overlap with basic word usage. So any attempt to date the phylogenetic origin of word-use will no doubt be challengeable on questions of where-to-draw-the-line as well as on the paucity of evidence. For the origins of drawings, we may look to evidence such as cave-paintings of animals dating back some 40,000 years; or even earlier Neanderthal cave art, e.g. at La Pasiega in Spain. For the written word we might look back some 5,000 years or more to Sumerian cuneiform used for keeping accounts.

This Third Draft explicitly incorporates the use of representations. Is there a 'symbol-grounding problem' yet? Is the relationship between words/symbols/representations and their referents at all problematic? Within the pragmatic cognitive framework so far presented, such issues do not seem present. But some people have concerns, so regrettably we must digress to mention them.

An aside: the Representation Wars

I am taking what I broadly call a pragmatic, dynamical systems approach to cognition. Clark (2001), from an opposing perspective, would characterise my position as asserting:

“Structured, symbolic, representational, and computational views of cognition are mistaken. Embodied cognition is best studied using noncomputational and nonrepresentational ideas and explanatory schemes, and especially the tools of dynamic systems theory.” (Clark, 2001, p. 128).

This is roughly correct — except that I strongly object to the characterisation of these two sides in the Representation Wars (Williams, 2018) as ‘representational’ and ‘nonrepresentational’ since it suggests the former take representations seriously and the latter have no use for them in their analysis. From my perspective the reverse is true. The former, in alluding to ‘internal representations in the brain’, are appealing to unanalysed, unexplained representations as *explanans*; whereas I see representations as *explananda* in need of explanation. A main theme of this paper is analysis of the development of representation-use, at increasing levels of complexity.

I have presented my arguments at length elsewhere (Harvey, 1996; 2005; 2008) and intend here to only summarise them as bullet points before returning to the main thrust of this paper.

1. Physicists will in some contexts use colliding billiard balls (and in other contexts waves) as a metaphor for the behaviour of subatomic particles. This is often useful.
2. Such metaphorical billiard balls are *explanans.* What is relevant is how they bounce off each other, but to ask how they themselves are materially constituted is to miss the point of the metaphor.
3. People explaining how complex systems (from central heating controllers to brains) typically use functional explanations that appeal to a homuncular metaphor (Harvey, 1996; 2008). Such a metaphor is often useful.
4. Within such explanations a component part (e.g. thermostatic switch, neuron) is treated as a little agent (homunculus) communicating via signals to other such metaphorical agents.
5. These metaphorical homunculi are *explanans*. What is relevant is their inter-agent relationship through communication, how they signal to each other ‘as if they were people’. They are, themselves, unexplained; this is appropriate for such metaphorical usage.
6. Whilst such functional explanations and homuncular metaphors are invaluable for all sorts of purposes, clearly the use of representation-using homunculi as unexplained *explanans* is a non-starter for explaining *how human cognition works*. A physicist cannot sensibly use a metaphorical billiard ball to explain the bouncing behaviour of a real billiard ball. That is *petitio principii.*

As a veteran of the Representation Wars, I am all too aware that these arguments usually fail to convince the so-called representationalists. The latest iteration of their viewpoint is Predictive Processing (Clark, 2016; Williams, 2018), that merely appeals to a different interpretation of what the metaphorical homunculi are doing. It puzzles me why such views persist, and I have two tentative suggestions. Firstly, the treatment of ‘internal representations in the brain’ without reference to them being relative-to-a-cogniser can be directly matched to the child who is too young to pass the Sally-Anne test (discussed above), who is unable to distinguish between Sally’s perspective and Anne’s perspective. The so-called representationalists are suffering a more sophisticated version of this infantile or autistic shortcoming, and they are simply unable to appreciate their own failings. Secondly, I notice that when challenged to provide an operational definition for identifying what will count as their ‘internal representations’, or their ‘predictions’, they typically do not understand why they should do so, they fail to deliver. This reinforces the impression that they are using representations as *explanans,* as unquestioned and unexplained postulates.

A further aside: the Extended fallacy

A further error associated with theories positing internal representations in the brain is that this implies that such representations can be identified with the state of affairs at a *location*. From my perspective, cognitive entities such as representations are necessarily relational and inter-personal. A signpost at a road junction does indeed have a location, but its role as representation (of direction to the next town) requires a linguistic community with a set of shared practices. It is a basic category error to suggest such cognitive entities have a physical location. As a corollary, discussion of ‘an extended mind’ (Clark and Chalmers, 1998) is equally absurd — unless the claim

is no more than the trivial observation that people can have cognitive relationships with something far away, like the sun, and they can make marks in the environment.

There is currently a movement (e.g. Newen et al., 2018) promoting '4E Cognition' — Embodied, Embedded, Extended, and Enactive; whereas 3 of these appear sound to me, the extended E is misguided and literally misplaced. The mistaken urge to locate cognitive entities such as representations in space is directly associated with the notion that they are objective and non-relational. But whereas for instance you may now be reading the written word 'ball' as a specific pattern located on a specific piece of paper or computer screen, it makes no more sense to identify that with the concept 'ball' than it does to identify a specific drop of water with a wave crossing the ocean. Mind and cognition does of course *relate* to things in the environment (Clark and Chalmers, 1998), as does the concept 'East'; that does not mean that mind and cognition *are* in the environment, no more than there is an East Pole located somewhere.

Does Life = Cognition? Draft 3 response

It is only after language becomes available to us that we can even ask Stewart's question: Does Life = Cognition? At this stage, people can competently discuss issues such as 'is this cow alive?' or 'does Sally know where I hid the sweet?', but most people will not understand the Life/Cognition question. For that you need to be a theoretician, to develop Theories of Life and Theories of Mind.

No doubt the first people to do so, historically, were religious theorists and philosophers: the scientists of their time. Historians of philosophy, from Aristotle on, can offer countless competing versions of Neurath's boat, navigating the cognitive ocean. Taking here an ontogenetic rather than historical perspective, we can say that the draft 3 response of a 21st century adult is likely to be: 'I don't understand the question. Sounds like we need some theory'.

And if representation-using is 'relationships-squared', then maybe here folk theories of cognition followed by scientific theories of cognition involves 'relationships-cubed', and to the fourth power… …, the manipulation of words in new dimensions.

1. The baby comes to relate to a sweet as something meaningful in its world. This is the basic relationship between person and thing that has significance for it.
2. "Where is the sweet?" says the mother to the baby, in repeated variant forms of interaction, until the baby starts pointing correctly in the basket and not pointing incorrectly elsewhere. "Well done !!" for this basic use of representations (both words and pointing). Perhaps 'relationships-squared': baby relates to {a word that relates to a thing}.
3. "Who knows 'where is the sweet?' ?" says the mother to the child taking the Sally-Anne test. When the child points correctly to Anne, and does not point incorrectly to Sally — "Well done !!!" for such basic folk cognition. Perhaps 'relationships-cubed': child relates to {a doll that relates to {a word that relates to a thing}}.
4. "Who knows 'who knows 'where is the sweet?' ?' ?" Baron-Cohen et al. (1985) have a theory, they correctly distinguish between the 3 year-old (or autistic) child who doesn't pass the Sally-Anne test, and the 4-year old that does. "Well done !!!!" for the scientific theory. Perhaps relationships to the fourth power: psychologists relate to {a child that relates to {a doll that relates to {a word that relates to a thing}}}.
5. "Who knows 'who knows' … …?'?'?'?'?" and we are into the realms of meta-theory — the theme of this paper.
6. … …

Our boat needs more superstructure.

Fourth Draft: an initial theoretical stance

When we draw lines on the sand to create a map, with pebbles to represent destinations or way-points, the symbols can play two different roles. On the one hand the map re-presents (presents again, in usefully equivalent form) the state of the world, the relationship between roads and

towns. On the other hand, the pebbles can be shifted, new lines added, to re-present hypothetical what-if scenarios; we start to theorise about the world. We can use symbols to talk-to-ourselves as well as to communicate with other people.

The pebble on the sand map is a real physical object located in space. The pattern of air-vibrations that we hear as the word ‘ball’ is a real physical pattern located in space and time. We can write the letters b-a-l-l on the sand, on paper, on the computer screen as real marks in space. What makes these into symbols rather than just physical things is the social context and learnt practices within which they affect behaviour. New kinds of generalisation become possible for us.

We already (Second Draft) learnt to generalise across different views of a physical ball, at different times, to achieve competency at object permanence. We learnt how to generalise across perspectives from different people, so we could pass the Sally-Anne test. At this stage (Fourth Draft) we start to generalise across all these different perspectives on a ball — the visual appearance and feel of an actual ball, the verbal and written instances of the word ‘ball’ — to form the concept of ‘ball’. We can count two balls, we can manipulate the symbol 2, we can even use it as part of the number 22.

An initial theoretical stance — and this is a characterisation of the position held by cognitivists and so-called representationalists — is to be impressed by the amazing powers that humans gain through use of representations, to realise how central this is to our cognitive world; and then to use such representations as the unanalysed *explanans* for their theories of cognition. This is one side in the Representation Wars.

The paradigmatic example of this is the interplay between the word ‘computer’ and the computationalist viewpoint. In the early 20th century ‘computers’ were people who worked in insurance offices and the like, processing quantities of data by hand in massive spreadsheets and performing manual calculations. Their symbols were pen on paper and their computations involved following algorithms such as that for long division. First mechanical devices were enlisted to aid them, then later in the 20th century electronic devices were developed took over all the tedious algorithm-crunching. And these devices also inherited the description ‘computers’. We could build machines that could replicate some important human capacities. Functional explanations held out the promise of explaining how brains work in these terms. GOFAI has arrived.

Does Life = Cognition? Draft 4 response

The sophisticated use of symbols such as words allows us to use concepts such as Life and Cognition. So someone at this level of competence may at least make an attempt at answering Stewart’s question: Does Life = Cognition? It sounds like Life is one thing, like pebble A, and Cognition is another thing, like pebble B. Are they one and the same thing — as when we learnt about object-permanence, or learnt that the Morning Star and the Evening Star were both the same planet Venus?

Parsed that way, the question allows for many possible answers, dependent on just what Theory of Life or Theory of Cognition/Mind you may subscribe to. Even the computationalist GOFAI position of Draft 4 could, I think, still endorse the response that the question so-parsed does not really make sense, since they are different category classes on each side of the equals-sign. Life and Cognition are not the same concept, but they do in some sense entail each other.

Fifth Draft: a noncomputational theoretical stance

We start by counting on our fingers, we initially learn our multiplication tables through chanting at primary school, we can eventually calculate the result of ‘2x22=?’ in our head. This does not mean that we are manipulating symbols that are actually ‘in our head’, or ‘in our brain’. But if that is not the explanation, what different explanation can be offered?

Synthesising the basic use of representations

One way of checking one's understanding of representation-using is to synthesise the behaviour in a robot and check carefully what assumptions you need to build in. Evolutionary Robotics, ER (Harvey et al., 1997; Harvey, 2005; Harvey et al., 2005) is one such approach. The synthesis of basic sensorimotor behaviour, such as using vision to move towards targets or to avoid some other class of object, is generally considered unproblematic in principle. However some consider that fresh difficulties arise with so-called representation-hungry problems (Clark and Toribio, 1994; Clark, 2001), as we progress from merely lower level sensorimotor engagement with the environment to higher level representational engagement.

But we have used ER to demonstrate in practice as well as in principle that we can synthesise the evolutionary origins of representation-using via sensorimotor patterns of robot behaviour in scenarios that may be simplistic, but are real physical ones (Quinn et al., 2003; Harvey, 2008). To demonstrate the basics of representation usage, we need several robots that are mobile and can sense (here with a form of vision) their environment and each other. In (Quinn et al., 2003), there are 3 such identical robots with short-range vision that were implicitly given the task of following each other in line across a plane; their fitness was based on their joint performance on this task (based on the distance travelled together), and this fitness influenced selection over successive generations. The evolutionary runs were performed first in simulations using simulated agents, but later similar results were obtained using evolution with real robots. Because of the limited vision, each robot was only capable of seeing one robot immediately ahead. The only way to achieve a moving line-of-three was for one robot to choose a Leader role, and the other two to choose Follower roles. They all started off in identical state, though randomly placed, and were provided with no further means of communication and no hint as to what Leader or Follower roles might entail. Nevertheless, if they were to choose non-conflicting roles, they would have to communicate somehow.

This scenario shows how ER can be used to address philosophical concerns (Harvey, 2000; 2005; Harvey et al., 2005). The human designers of the experiments do not instruct the robots to behave in a desired manner, since the construction of the robots and the design of the fitness function and of the (artificial) evolutionary protocol requires no reference to any behavioural criteria. Despite this, robot fitness will in practice increase if the robots do behave as the experimenters wish. Like proud parents, the experimenters will be pleased if the robots 'figure it out for themselves' without even 'being told what the desired behaviour is'.

In these experiments the robots did indeed succeed in communicating, through using their sensorimotor interactions. We can provide a post-hoc analysis of how they managed to do it, at a behavioural level of description. It looks like they each started moving in a pattern so as to locate the others and form a cluster. Some form of symmetry-breaking was then needed, to differentiate one as Leader and the others as Followers. It looks like this symmetry was broken by whichever happened to be the first to make some stereotypical movement within sight of another, that triggered reactions so as to affect the joint dynamics of all three. This analysis is at a behavioural level and described anthropomorphically ('…it looks like…'). We also have, because this is ER, a complete description of the evolved control system ('brain and body') that provides a parallel mechanism level of explanation that does not use such behavioural language.

That stereotypical movement — we are not fully sure what it was, or what parts of the evolutionary history that it needed to arise — *represented* the message that might be translated 'I am the Leader, you become Followers'. We have a signal, a sender and two receivers, and a context in which the message is meaningful. This is of course at the very simplest end of the scale of representation-hungry problems; but it shows an ER approach to synthesising representation-using in robots that in principle could be extended to more sophisticated representation-use.

It should be noted that the signal, the representation, in this ER example is not a discrete symbolic token. It is the conjunction of a real-life movement of flesh and blood — well, maybe that should be rephrased as a real-robot movement of metal-and-motor — with a social (multi-robot) embodied sensorimotor context wherein appropriate responses are triggered.

Though here I have not used the language of enaction, the views are in the same ballpark as enactivists (Stewart, 2010). Likewise there are clear relationships with those who relate language to embodiment (Lakoff and Johnson, 1980; Di Paolo et al., 2018).

Does Life = Cognition? Draft 5 response

We still have Life not equal to Cognition, but we now have a clearer view of just how cognition — and specifically here the use of representations — can arise from an embodied life-form or its simplified robotic part-analogue. We can see how such cognitive competencies might be explained without appeal to magic. But how about Life?

Sixth Draft: a look at Life

So far the focus has been on the Cognition part of the equation. Life has been taken as an unanalysed given. I have not yet made any use of ideas arising from autopoiesis (Varela et al., 1974). I would hope that what I have said so far is compatible with an enactive perspective, and I consider my appeal to the Neurath's boat metaphor to be very much in the same spirit. But I simply have not yet needed any appeal to technical aspects of autopoiesis in order to make my points. As far as I can see, they fit comfortably with pre-theoretical ideas of folk biology and folk cognition, and with 'Fourth and Fifth Draft' levels of competence. In these respects at least, a more scientific analysis has not yet required any serious rebuilding of the boat before reaching these conclusions.

But if cognition is, roughly, what life does and rocks do not, then we may want to analyse the difference between living things and rocky things. In particular they both appear to be constituted out of what physicists call matter, so unless we are going to appeal to some mysterious 'life force' it looks like the difference is in how the matter is organised.

Autopoiesis takes this latter choice (Varela et al., 1974), and proposes a method for the scientist-observer to carve up the natural world of matter at some convenient joints, that distinguish individual living organisms from the environment they inhabit. Here I am not going to tie the Fifth Draft of my Neurath's boat to the specifics of autopoietic theory, but instead sketch a weaker version of Autopoiesis-Lite. The -Lite suffix here indicates that this will cover a broader category than just living organisms, it will cover cyclones as well.

Whirlpools and cyclones are familiar examples of physical systems that can in the right conditions maintain a pattern of dynamics within some local region. They are driven by environmental energy gradients, they are constituted by a flow of matter, such that though the specific matter is transient in the short term it is the pattern of dynamics that persists over a longer term. They persist despite some external perturbations, yet may be vulnerable to other perturbations that make them disappear; so the language of life and death for such processes fits easily. They are individuated entities, distinguishable from their environment and from each other; hence we can give cyclones names, e.g. Hurricane Katrina. Since their persistence and their individuation are not imposed by external agencies, they are often called self-maintaining and self-individuating. This self- prefix may be unfortunate and misleading in implying the agency is internal, when in fact it is the interplay of both internal and external dynamics that explains these phenomena.

Autopoietic theory (Varela et al., 1974) goes much further than this, in terms of formal specification and requirements for the maintenance of the boundary; it will exclude cyclones from its remit. For the purposes of this paper it is sufficient to go with this broader 'Lite' version.

Behavioural descriptions and physical descriptions

Autopoiesis-Lite approaches are inspired by such physical systems which are seen to epitomise some basic important universal characteristics of living organisms. Such a perspective defuses in-principle concerns about the Origin of Life, since the origin of e.g. a cyclone does not raise such concerns. It allows the observer to check what external or internal factors *threaten* the persistence of the dynamical pattern, and what *regulatory* responses to these *threats* tend to

prolong its persistence. In fact this perspective allows us to start using these cognitive and motivational terms such as *threat* and *regulatory* in what can nevertheless still also be fully described in the language of physics.

Cyclones and whirlpools, though sharing crucial features with living organisms, are somehow too boring for us to call them alive. Let us provisionally call them proto-organisms. One thing they significantly lack is any significant history arising from their interactions with their environment. They do not appear to learn to learn much from their experience. Although Hurricane Katrina is sufficiently individuated to justify carrying a name, this is individuation by location rather than individuation by distinctive features. Cyclones and whirlpools carry no evolutionary history with them either through any internal equivalent of DNA or via external traces left in the environment.

We should mention here Cairns-Smith (1985) who proposes that it is exactly those sort of features, the building up of historical distinctions, that might explain the historical steps from whirlpool-like entities to possible early organisms. His suggestion is that the normal physical processes of clay crystals interacting with the flow of a stream can result in differential survival rates of variant crystals; and such clay crystals can replicate so as to preserve variant features in the copies. Thus, Cairns-Smith suggests, basic non-physical evolution could bootstrap Darwinian evolution of more biological organisms.

For an Autopoiesis-Lite perspective on the relationship between a living organism and the matter it is made of, we do not actually need any Darwinian underpinning. But it would seem that any self-maintaining individual needs some specific history attached to it (whether through DNA-equivalent or otherwise) for it to be plausibly any more than just a proto-organism.

Though cyclones lack this, even they allow observers of such phenomena to legitimately talk of what things or processes in the environment are of concern to a cyclone. Hume (1739) posed the Is-Ought problem: how can we coherently move from descriptive statements to prescriptive statements, from an Is to an Ought. There is the equivalent issue here, how do you coherently get from neutral descriptive statements to the language of concerns such as ‘X is a *threat* to Y’? Autopoiesis, even in an Autopoiesis-Lite version and with something as relatively simple as a cyclone, provides that justification. An individual is defined, carved out of the world, and we can now discuss the behaviour of this individual-as-a-whole. Its interactions with its world around it can be described in terms of needs, concerns, threats, regulation; ultimately grounded in its prospects for continued survival or its ceasing to exist, its ‘death’, but further contingent concerns arising from this. We can discuss its perception of the world about it; not only how we the observer can conveniently carve out individual A from the rest of *our* world, but also how individual A can carve up *its* world in ways that make sense for it. This is also Enaction-Lite.

Does Life = Cognition? Draft 6 response

We have already, in previous Drafts, used behavioural language as well as the language of physics. A major step forward with Draft 6 is that— even with something as simple as Autopoiesis -Lite — we start to appreciate just how these two languages relate to each other. The medical doctor can use either language in dealing with her patient: the physical (‘taking this medicine X will result in Y’) and the behavioural (‘if you want to get healthy I encourage you to take X regularly’).

Does this change any answer to Stewart’s question: Does Life = Cognition? My answer will still be ‘no, they are definitely not the same concept, but they do in some sense entail each other’. But now the nature of that entailment is somewhat clearer. Autopoiesis-Lite clarifies the relationship between living organisms (even simplistic proto-organisms like cyclones) and physics. The nature of that relationship explains how we can use behavioural language, and explains the difference between a blink and a wink. The blink is a nervous tic, not a behaviour; but a wink is a behaviour, used by one individual to communicate to another, to convey meaning. Cognition is ultimately the study of such behaviours.

So at this Draft 6 level of understanding of cognition, Life, or something equivalent, is the prerequisite for using the language of behaviour, which is what we require to study Cognition. The qualification ‘or something equivalent’ is partly there because Autopoiesis-Lite also covers proto-

organisms (e.g. whirlpools and cyclones) that we probably do not want to call alive. We can and do nevertheless extend behavioural language to these, in at minimum a metaphorical fashion.

One may further (and probably controversially) argue that with ER, evolved robots partially satisfy the conditions of Autopoiesis-Lite: the physical robot does not individuate itself, but the evolved behavioural habits that survive over generations do so, within the experimental context. Less controversially, an unevolved robot, whose control system may have been hand-crafted by engineers, nevertheless lends itself to a behavioural level of description. Even if this is parasitic on the designers' intentions, we can still meaningfully discuss the cognition of such robots. As-if alive robots lend themselves to as-if cognitive descriptions.

This in turn supports the use of homuncular metaphor in functional analyses of complex systems, e.g. the brain. But such functional analyses of the brain cannot possibly, cannot as a matter of logic, explain cognition. We have argued above that the cognitivist attempts to explain cognition in terms of cognition are unprincipled and comparable to an infantile failure to pass the Sally-Anne test. Autopoiesis, or as here Autopoiesis-Lite, provides a principled approach.

So far the progression through 6 drafts of cognitive perspectives have led to a sketchy description of my own position. Though many who follow enactivist or autopoietic approaches will no doubt disagree with some specific points, I think they will be broadly sympathetic. But we can now make a further move that may leave well them behind. A major plank, perhaps the keel of Neurath's boat, is rotten and needs replacing.

Seventh Draft: Mind from Matter, Matter from Stuff

Previous shifts from one level of cognitive understanding to another new level have typically involved expanding a simplistic objective view of some entity into a more sophisticated relational interpretation. For example:

1. Before recognition of object-permanence, 'object-exists' equated to 'object-is-visible', implying that if it goes behind a tree it ceases to exist. With competence at understanding object-permanence, a child starts to recognise that what-is-visible varies according to the relative positions of child, object and potential visual shields like trees.
2. Before a child passes the Sally-Anne test, the child assumes that what-Sally-knows is the same as what-Anne-knows. Afterwards the child starts to realise that such knowledge is not objective and universal, but is relative-to-a-knower.
3. The computationalist theorises on the basis that neurons in the brain (or a pattern of neural activation) can act as some Platonic form of representing symbol. The transition above from Draft 4 to Draft 5 comes from appreciating the relational nature of representations embedded in a social context, Platonic symbols just do not work.

I argue (Harvey, 2000) that science requires objectivity; but that major advances in science have arisen from recognising where there is observer-dependence of phenomena that can be fashioned into a higher level of inter-subjective agreement. In astronomy, the Copernican revolution abandoned our privileged Platonic position at the centre of the universe and instead asked how the solar system would look viewed from the Sun or another planet. The relativism of Copernicus extended the realm of the objective by acknowledging our subjective viewpoints. Einstein carried the Copernican revolution further with Special Relativity, by considering the viewpoints of observers travelling near to the speed of light and insisting that scientific objectivity required their perspectives to be equally privileged to ours. These scientific advances are sophisticated versions of passing new Sally-Anne tests.

Above we have sketched out how a world of matter can be seen as dividing into Life (organisms made of matter) and its Environment (also made of matter); how the persistent though fragile forms of life underpin our ability to describe their actions in behavioural terms. We have the beginnings of a theory of Life and a theory of Cognition. But these theories are presented as our scientific view-from-outside, as our objective description of a world with cognitive beings. I now suggest that we have failed to distinguish between matter as we the external observers

understand and experience it, and matter as the cognitive being in our theory understands and experiences it. A major plank of the boat is perhaps rotten.

Different worlds

In robotics, including ER, the novice roboticist fairly quickly comes to appreciate that what is obvious to the human engineer, e,g, the door in the wall, is not at all obvious to the robot. More than a century ago a couple of literary works (Abbott, 1884; Hinton, 1907) highlighted this issue in what nowadays might be called Artificial Life speculations about Flatland. This is 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. From our 3D perspective, we as external observers can imagine a sphere passing through this plane, and note that to the inhabitants of Flatland (who just experience its intersection with their world) it would appear as a dot out of nowhere, expanding to a circle that then contracts and disappears. The message to take away (Harvey, 2019) is that the objects we observers relate to (that include spheres) are not, indeed cannot be, the same objects that the agents experience.

We can make coherent sense of this, without any contradictions, if we relativise experienced objects as either objects-for-the-external-observer, or objects-for-the-Flatland-inhabitants and make sure that we distinguish between them. The external observer experiences the former objects first-hand, but experiences the latter objects only second-hand via her appreciation of how the Flatland inhabitants experience them. They cannot be equated. This does not raise problems when we humans analyse cognition in Flatland, but does when we try to analyse cognition in our own human world (Harvey, 2019).

Matter and Stuff

Throughout this paper, through successive drafts of theories of cognition, we as scientists have been discussing living organisms made of matter, and their interactions with their environment made of matter. But if those organisms are us, we have a contradiction. We must distinguish between matter-for-humans-we-theorise about and matter-for-us-the-theorists. Copernicus and Einstein had to be careful introducing relativism into astronomy and physics, and relativist cognition is at least as tricky. Can we resolve this, perhaps in some way analogous to coping with the Sally-Anne test?

A speculative proposal hinted at in Harvey (2019) is to postulate some more objective replacement for physical matter that we shall call Stuff. For many purposes Stuff and physical matter can initially be treated as identical. We now rephrase our theories of Life and Cognition in terms of Stuff rather than matter. Autopoiesis, or Autopoiesis-Lite, now refers to the way that physical interactions of a world of Stuff allow us to carve out self-individuated and self-maintaining processes that we identify with organisms or proto-organisms. We are now able to describe the interactions of these organisms with their environment in behavioural terms; we have a grounding for a theory of cognition.

So far, we have achieved nothing new by just renaming matter as Stuff. But now we should start to appreciate that any theory of physics developed by people will be grounded in how they experience their world. If they call that basis of their physics Matter (capitalised to note that it is matter-for-them) then Matter and Stuff cannot be identical; cf. the Flatlanders. What makes it different from the Flatland scenario is that, here, talk of ‘how *they* experience *their* world’ is actually talk of ‘how *we* experience *our* world’ (Harvey, 2019).

The normal failure to make any distinction between Matter and Stuff may go some way towards explaining the unease many people have in looking at the relationship between Life and Matter (or indeed subjective consciousness and Matter). The Matter of physics is essentially dead (and lacking in subjective experience); so how can it possibly support the emergence of Life (or subjective experience)? One way to address this concern is to rephrase this in terms of it being some other Stuff, not Matter, that supports the emergence of Life (and consciousness). Matter is how we experience Stuff.

This sounds like a fanciful speculation that achieves rather little. But *if* we want to propose some overarching combined theories of Life, Cognition and Mind, built from a single substrate, the arguments I have presented suggest that substrate cannot be the Matter of our physical theories, but rather some other substrate that I have called Stuff. I accept that many will not want to go down that road, but the price of that choice would be abandoning the hope of such overarching combined theories based on Matter.

Conclusions

The context for this paper was Stewart's (1992, 1996) suggestion that Life = Cognition. We cannot address the truth or falsehood of this without discussing the meaning of these terms, and the relationship between them. We need theories of life, and theories of cognition, and I argue that these cannot be generated from a disembodied and unsituated armchair. So, starting from where we are, I presented a trajectory of cognition developing through infancy, through pre-theoretical and theoretical stages, culminating in more or less sophisticated theories of life and cognition. The metaphor of Neurath's boat was used throughout: each new stage is not built from scratch, it arises from a reworking of flaws in the earlier stages becoming visible. The Sally-Anne test characterised one early stage in infancy, and several later rebuilds of the cognitive boat could be interpreted as dealing with higher level versions of this Sally-Anne test.

There is an inevitable circularity in the use here of Neurath's boat: we have to describe the trajectory of boat rebuilding from a perspective biassed by our current position in that trajectory. The final speculation on Matter and Stuff reflects the complexities of circular causation (Harvey, 2019).

I generally find myself in broad agreement with Stewart's (1992; 1996; 2010) approach to these issues whilst disagreeing with him on some specifics. My conclusions on the relationship between Life and Cognition probably fits that pattern: my Autopoiesis-Lite and Enaction-Lite place me broadly in the same camp as the heavyweights, without signing up to all the specifics. I conclude that Life and Cognition clearly should not be equated, but that the development of a theory of life and a theory of cognition clarifies the entailment relationship between them. Autopoesis, even in its Lite version, sketches out how the world can be carved into dissipative self-maintaining structures that are individuated; cyclones still count here, though with their absence of accumulated history such proto-life is far less interesting than real life. This underwrites the move from a merely physical description of the dynamics to a behavioural language describing what concerns the individuated (proto-) life form. We can now talk of winks as well as blinks, the study of cognition becomes possible.

And the study of cognition is really hard. Neurath's boat has some doubtful woodwork, the seas are rough.

References

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

Baron-Cohen, S., Leslie, A. M. and Frith, U. (1985). Does the autistic child have a "theory of mind"?. *Cognition* 21(1): 37–46.

Bourgine, P. and Stewart, J. (2004). Autopoiesis and cognition. *Artificial Life* 10: 327-345.

Cairns-Smith, A. G. (1985). *Seven clues to the origin of life*. Cambridge University Press.

Clark, A., & Toribio, J. (1994). Doing Without Representing? *Synthese* 101: 401- 431.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis* 58(1): 7-19.

Clark, A. (2001). *Mindware*. New York and Oxford: Oxford University Press.

Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.

Di Paolo, E. A., Cuffari, E. C., and De Jaegher, H. (2018). *Linguistic Bodies: The Continuity between Life and Language*. MIT Press.

Harvey, I. (1996): Untimed and misrepresented: connectionism and the computer metaphor *AISB Quarterly*, no. 96, pp. 20-27.

Harvey, I., Husbands, P., Cliff, D., Thompson, A. and Jakobi, N. (1997): Evolutionary robotics: the Sussex approach. *Robotics and Autonomous Systems*, v. 20, pp. 205–224.

Harvey, I. (2000): Robotics: Philosophy of Mind using a Screwdriver
In *Evolutionary Robotics: From Intelligent Robots to Artificial Life, Vol. III*, T. Gomi (ed), AAI Books, Ontario, Canada, 2000. pp. 207-230. ISBN 0-9698872-3-X.

Harvey, I., (2005): Evolution and the Origins of the Rational. In: Zilhão, António (ed.), *Cognition, Evolution, and Rationality*. London, Routledge, 2005. Routledge Studies in the Philosophy of Science. ISBN 0415362601.

Harvey, I., Di Paolo, E., Wood, R., Quinn, M, and E. A., Tuci, (2005). Evolutionary robotics: A new scientific tool for studying cognition. *Artificial Life*, 11(1-2), pp. 79-98.

Harvey, I. (2008). Misrepresentations. In S. Bullock, J. Noble, R. A. Watson, and M. A. Bedau (Eds.) *Proceedings of the Eleventh International Conference on Artificial Life*, pp.227-233, MIT Press, Cambridge, MA. ISBN: 978-0-262-75017-2

Harvey, I. (2019). Circular causation, circular cognition: a tour around common confusions. Accepted for special issue of *Artificial Life*, in press.

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

Hume, D. (1739) *A Treatise of Human Nature*. London: John Noon, at the White-Hart, near Mercer's-Chapel, in Cheapside.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Neurath, O. (1921). Anti-Spengler. Vienna Circle Collection vol. 1: *Empiricism and Sociology* (1973). doi:10.1007/978-94-010-2525-6_6. ISBN 978-90-277-0259-3.

Neurath, O. (1944). Foundations of the Social Sciences. In O. Neurath, R. Carnap, Ch. Morris (eds.), *International Encyclopedia of Unified Science*, vol. 2, n.1, Chicago: University of Chicago Press.

Newen, A., De Bruin, L. and Gallagher, S. (2018). *The Oxford Handbook of 4E Cognition*. Oxford University Press. ISBN: 9780198735410. DOI: 10.1093/oxfordhb/9780198735410.001.0001

Piaget, J. (1977). Gruber, H. E. and Vonèche, J. J. (eds.), *The essential Piaget*. London: Routledge and K. Paul.

Quine, W. V, (1950). Identity, ostention and hypostasis. *The Journal of Philosophy*, 47(22): 621-633.

Quine, W. V. (1960). *Word and object*. MIT Press.

Quinn, M., Smith, L., Mayley, G. and Husbands, P. (2003). Evolving controllers for a homogeneous system of physical robots: Structured cooperation with minimal sensors. *Philosophical Transactions of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, 361:2321-2344.

Steele, M. A., Halkin, S. L., Smallwood, P. D., McKenna, T. J., Mitsopoulos, K., Beam, M. (2008). Cache protection strategies of a scatter-hoarding rodent: do tree squirrels engage in behavioural deception? *Animal Behaviour* 75(2): 705-714

Stewart, J. (1992). Life = cognition: the epistemological and ontological significance of artificial life. In: Varela, F. J. and Bourgine, P. (Eds.), *Toward a practice of autonomous systems: Proceedings of first European conference on artificial life*. ECAL1991, pages 475-483. MIT Press.

Stewart, J. (1996). Cognition = life: implications for higher level cognition. *Behavioural Processes* 35: 311-326.

Stewart, J. (2010). Foundational Issues in Enaction as a Paradigm for Cognitive Science: From the Origin of Life to Consciousness and Writing. In Stewart, J., Gapenne, O. and Di Paolo, E. E. *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press. DOI:10.7551/mitpress/9780262014601.003.0002

Varela, F. J. and Bourgine, P. (1992), *Toward a practice of autonomous systems: Proceedings of first European conference on artificial life*. ECAL1991. MIT Press.

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.

Williams, D. (2018). Predictive Processing and the Representation Wars. *Minds and Machines,* 28(1): 141–172.