

Friends Reunited? Evolutionary Robotics and Representational Explanation

Michael Wheeler

Department of Philosophy

University of Dundee

Dundee, DD1 4HN, Scotland

email: m.w.wheeler@dundee.ac.uk

telephone: +44 1382 344517

Abstract

Robotics as practised within the artificial life community is no longer the bitter enemy of representational explanation in the way that it sometimes seemed to be in the heady, revolutionary days of the 1990s. This rapprochement is, however, fragile, because the field of evolutionary robotics continues to pose two important challenges to the idea that real-time intelligent action must or should be explained by appeal to inner representations. The first of these challenges, the threat from non-trivial causal spread, occurs when extra-neural factors account for the kind of adaptive richness and flexibility normally associated with representation-based control. The second, the threat from continuous reciprocal causation, occurs when the causal contributions made by the systemic components collectively responsible for behaviour-generation are massively context-sensitive and variable over time. I argue that while the threat from non-trivial causal spread can be resisted, the threat from

continuous reciprocal causation provides a stern test for our representational intuitions.

Keywords: evolutionary robotics, intelligent action, modularity, representation.

1 An Unfinished Story

As soon as one starts to think in general terms about ALife-related robotics (by which I mean nothing more than robotics research which is standardly reported in artificial life journals and at artificial life conferences), it is very likely that the following narrative will come to mind. Recent attempts to explain and to replicate artificially the ways in which humans and other animals produce fluid and flexible, real-time adaptive responses to ongoing sensory stimuli have come to reject what might fairly be called the orthodox view in artificial intelligence. According to this view, real-time intelligent action of the kind just described is possible only because the agent has the capacity to build and to manipulate detailed internal representations of the external world. But however intuitively plausible the orthodox approach may be, and however influential it has been (and it scores highly on both points), it faces a serious problem. The demand that complex inner representational structures be constructed and maintained in real time, in the face of dynamic, noisy, and sometimes unforgiving environments, establishes an information-processing bottleneck in the perception-action cycle. This bottleneck constitutes an obstacle to adaptive success that may well be impossible to overcome, given biologically plausible processing resources. So what is to be done? The response, within ALife-related robotics, has been to develop robot control systems that side-step the apparently troublesome need for computationally expensive inner complexity by relying on frequent sensing of the environment rather than the building of internal world models. This, of course, is the now-familiar move towards what Brooks has dubbed ‘situatedness’ (see, e.g., [8]).

At this point in our narrative, we need to pause. If this paper were being written in the mid-1990s, the next plot development would no doubt have been the explosive suggestion that

the explanation and replication of real-time intelligent action might profitably do away with the appeal to inner representation altogether (see, e.g., [3, 20, 37, 38]). But now, just a few years on, the introduction of this radical assault on our most cherished of explanatory concepts would no longer seem to reflect the ALife zeitgeist in the way that it once did. Somewhere along the line, peace seems to have broken out between the ALife-related roboticists and the advocates of representational explanation. As Randy Beer and Inman Harvey have pointed out to me (in conversation), it may be that the apparent rapprochement here is illusory, and that what has actually happened is that the two sides in the debate have simply bored each other into silence. However, I am still inclined to think that something like a reconciliation has occurred. But how? The answer is that early on in ALife-related robotics, a transformation in the very notion of representation took place, and this reconceived understanding of representation is not at odds with, indeed it may be a crucial weapon within, the situated approach to real-time intelligent action.

So what happened? According to the standard view in artificial intelligence and cognitive science, representations are conceived as essentially objective, context-independent, action-neutral, stored descriptions of the environment. These descriptions are built during perception, and then accessed and manipulated downstream by centrally located reasoning algorithms that decide on the best thing to do, in order to achieve certain current goals. By contrast, many ALife-related roboticists (as well as some recent researchers in computer vision) have come to adopt a view of representation according to which the states concerned are egocentric rather than objective (e.g., spatial maps built in agent-centred co-ordinate systems), transient rather than stored and recalled, and context-dependent rather than context-independent. Moreover, these states function not as action-neutral descriptions of the environment, but more directly

as control structures for actions. Indeed, as I have argued elsewhere [42], one might reasonably say that, according to the transformed concept of representation, how the world is should itself be encoded in terms of specifications for possible actions (something like Gibsonian affordances [19]). States with the conceptual profile just identified may illuminatingly be called *action-oriented representations*. (I have taken the term from Clark [9, pp.47-51]. However, it should be noted that Clark stops short of the full neo-Gibsonian position, preferring to say that action-oriented representations are poised between the dual functions of mirroring reality and controlling action, in the sense that the representations concerned emerge as being both encodings of how the world is and, simultaneously, specifications for appropriate actions. For discussion of this theoretical difference, see [42]. For seminal examples of action-oriented representation at work, see [1, 2, 8, 17, 29].)

It might seem, then, that with the concept of action-oriented representation, harmony has been restored. But perhaps we have spoken too soon. For there is at least one area of AI-related robotics which continues to supply a mandate for a kind of principled resistance to the peace process. That area is *evolutionary robotics*, the sub-discipline of AI in which algorithms inspired largely by Darwinian evolution are used to automatically design the control systems for (real or simulated) artificial autonomous agents. (For useful points of entry to the field, see [24, 31].) Roughly speaking, the evolutionary robotics methodology is to set up a way of encoding robot control systems as genotypes, and then, starting with a randomly generated population of controllers, and some evaluation task, to implement a selection cycle such that more successful controllers have a proportionally higher opportunity to contribute genetic material to subsequent generations, i.e., to be ‘parents’. Genetic operators analogous to recombination and mutation in natural reproduction are applied to the parental genotypes to produce ‘children’,

and (typically) a number of existing members of the population are discarded so that the population size remains constant. Each robot in the resulting new population is then evaluated, and the process starts all over again. Over successive generations, better performing controllers are discovered.

So why does evolutionary robotics provide evidence that real-time intelligent action may often be produced by non-representational means? The crucial observation here is that certain key examples of evolved robot control systems exhibit two intriguing phenomena which, as we shall see, make trouble for representational explanation. The first is an effect that Clark and I have dubbed *non-trivial causal spread* [43]. The second is an effect that, following Clark, I shall call *continuous reciprocal causation* [9]. To be clear: there is no *a priori* reason why human robot builders couldn't develop control systems which harness these apparently representation-threatening phenomena. Indeed, in the case of non-trivial causal spread at least, there are many examples in the ALife-related robotics literature where that has been done. However, adaptive solutions which exploit these phenomena do not present themselves to the human designer as naturally or as easily as do other strategies. (More on this later.) As any ALifer will tell you, it's received wisdom that artificial evolution (like its natural counterpart) has an uncanny knack of producing powerful and efficient solutions to adaptive problems, solutions which the vast majority of human designers wouldn't even have contemplated. So it is perhaps unsurprising that evolutionary robotics provides a fertile breeding ground for behaviour-generating strategies that involve non-trivial causal spread and continuous reciprocal causation.

The goal of this paper is to investigate these two threats to representation, as they emerge in evolutionary robotics research. The global anti-representational terrain here, and even the local area of that terrain occupied by evolutionary robotics, are territories through which I've

travelled before, sometimes as a solo explorer [38, 39, 40, 41] and sometimes (in the case of non-trivial causal spread, although not in the specific context of evolutionary robotics research) in partnership with Andy Clark [11, 43]. The present treatment builds on my prior investigations, in that various ideas and arguments to be found in that existing body of work have here been co-opted, adapted, modified and supplemented, as I now feel to be necessary. Much of what follows is based on material from [42].

Before we get down to business, I need to engage in some pre-emptive clarifications, prompted by the helpful comments of an anonymous referee. Given what I've said so far, one might expect my next step to be a general account of what it is for a state, process or system to be representational in character. But it won't. My current best shot at the necessary and sufficient conditions for representation (as that notion matters to cognitive science) will emerge in full as the argument of this paper unfolds. All I wish to assume at the outset are the (I trust) anodyne points (a) that the concept of representation is linked conceptually to the notion of an *information-bearing* resource, and (b) that the idea of a representation is at least the idea of an entity that *stands in*, within some processing economy, for some other entity or state of affairs.

On a different clarificationary note, the goal of this paper is to understand if, and if so how and under what conditions, the phenomena of non-trivial causal spread and continuous reciprocal causation might undermine representational explanation. But even if it turns out that the representationalist has genuine cause for concern here, the danger will be contained in two important senses. First, I have been careful to stage the present possibility of an anti-representational challenge exclusively within the domain of real-time intelligent action (i.e., the production of fluid and flexible, real-time adaptive responses to ongoing sensory stimuli). Large areas of what is standardly thought of as higher-level cognition, areas such as concept learning,

reflective advance planning, and the inner mental rehearsal of past or possible events, are just not in this category. So this paper has nothing to say about them. As it happens I have noted elsewhere [42] that there seems to be no principled reason to think that the scope of continuous reciprocal causation (in particular) is necessarily restricted to the domain of real-time intelligent action. However, in my view, it is only in that domain that the possible presence and power of this phenomenon have been demonstrated beyond doubt. So it is to that context that, at present, any anti-representational fallout from a recognition of continuous reciprocal causation ought to be restricted. In any case, there may be good independent reasons to think that representations of an essentially conventional kind will help to illuminate other kinds of cognitive achievement, such as those 'higher-level' capacities that I have already mentioned [42]. The second containing influence on any anti-representational scepticism that might, in the end, be mandated by what follows turns on the fact that this paper aims to use philosophical reflection on evolutionary robotics in order to identify certain conditions under which representational explanation might break down. It does not aim to pronounce on the extent to which such conditions will ultimately be found to exist in nature. That is a matter to be decided by empirical investigation, not armchair reflection.

2 Spreading Scepticism

As I shall use the term, *causal spread* obtains when some phenomenon of interest turns out to depend, in unexpected ways, upon causal factors external to the system previously or intuitively thought responsible [43]. To get a preliminary grip on the effect, as it occurs in the kind of context in which we are interested, consider a simple but compelling example which I have

adapted, with local variations, from John Haugeland [22]. (The ‘local variations’ mean that a truly excellent gag gets lost. I recommend checking out the original.) One way to succeed in driving from Edinburgh to Dundee would be to consult a cognitive map of the route, that is, to access a stored inner representation of how to get from the former to the latter. An alternative method would be to select the correct road in Edinburgh, and then to follow the signs until you arrive in Dundee. In the second case, the driver’s psychological innards and the road *collaborate as partners* in the successful completion of the activity. This partnership needs to be understood a certain way. It is not merely that the environment is a cheap and up-to-date source of information (although it is that), but that any adequate characterization of the intelligence-producing mechanism at work here would plausibly need to value the contribution of the road as being similar to that of the inner representation cited in our first solution. Thus, as Haugeland puts it, “much as an internal map or program, learned and stored in memory, would... have to be deemed *part of* an intelligent system that used it to get to [Dundee], so... the *road* should be considered *integral* to [the] ability” [22, p.235, original emphasis]. Although Haugeland doesn’t use our chosen term, this is an illustration of causal spread. Our initial expectation that the source of the observed intelligence must reside entirely inside the agent’s head (i.e., in a cognitive map) is replaced by an explanation in which an environmental factor (the road) makes a contribution which is arguably equal to that of any inner state or inner processes.

From the definition of causal spread given at the beginning of this section, taken in conjunction with Haugeland’s example, it should be obvious that to identify a case of causal spread is not simply to observe of some system of interest that that system is able to function adequately only when it is in interaction with some appropriate sort of environment. Rather, the *spread* in

question occurs with respect to some previously entrenched understanding of where the causal factors which explain the phenomenon of interest are spatially located, which is why an element of surprise is involved. So the observation that walking requires a supporting surface does not indicate a case of causal spread.

To make further progress, we need to take on board two additional points, namely (i) that causal spread comes in trivial and non-trivial forms, and (ii) that it is only the latter form which might even conceivably lead us in the direction of scepticism about representation. To unpack these points, we can begin with a methodological observation to which we shall return later: the core explanatory job that we have traditionally expected the concept of representation to do, *as a theoretical term in cognitive science*, is to help us understand the kind of contribution that neural structures and events make to the production of psychological phenomena and, in particular, to the generation of intelligent behaviour. In other words, we think that neural states and processes do something distinctive, and we expect the concept of representation to help us articulate what that something is, as well as how it occurs. Now recall that, in the present action-oriented climate, the idea that representations are objective descriptions of the environment has been replaced by the idea that representations are context-dependent codings for actions. What this tells us is that the claim which the recognition of causal spread would need to undermine, in order to generate a case against neural representations, is that neural factors constitute such codings.

One thing is immediately clear. To undercut the highlighted claim, it is not enough merely to demonstrate the existence of causal spread in the context of real-time intelligent action, i.e., to demonstrate merely that extra-neural factors unexpectedly contribute in important ways to such behavioural success. To see why, imagine a bizarre universe in which laptop computers occurred

in the wild, alongside trees and flowers. Imagine further that each of these naturally occurring laptops is capable of producing a range of behaviours, and that (fortuitously for the biologists of this world) a copy of the high-level source code that controls each of these behaviours is attached to each computer. For years, biologists interested in the outputs of these strange organisms concentrate all their attention on understanding the batches of representational instructions (codings for actions) given as high-level source code. Then, one day, it is discovered that certain additional features, such as a source code interpreter (e.g. a working compiler), are necessary for any output to ensue. This discovery would plausibly be a case of causal spread. Nevertheless, the addition of these particular causal factors to our explanation doesn't seem to have any adverse consequences for the entrenched understanding of the high-level programs as being representational in character. So mere causal spread is not enough to drive an anti-representational scepticism.

Time now to consider *non-trivial causal spread* [43]. This phenomenon arises when the newly discovered additional causal factors reveal themselves to be *at the root of some distinctive target feature* of the phenomenon of interest. For instance, where the phenomenon of interest is real-time intelligent action, one might single out the adaptive richness and flexibility of such behaviour as just such a feature. Haugeland's navigational example presents a clear case of such non-trivial causal spread. However, it's the kind of intuitive example that some people regard with suspicion. So here is a less contentious, experimental result in which non-trivial causal spread is present. This result is taken from evolutionary robotics.

Consider the following problem. A robot with a control system comprising an artificial neural network and some rather basic visual receptors is placed in a rectangular dark-walled arena. This arena features a white triangle and a white rectangle mounted on one wall. Harvey

et al. [21] gave artificial evolution the task of setting up the robot's control system so that, under wildly varying lighting conditions, it would approach the triangle but not the rectangle. The specific architecture of the neural network, the way in which the network is coupled to the visual receptors, and the field-sizes and spatial positions (within predetermined ranges) of those visual receptors were placed under evolutionary control. The result was a canny and genuinely unexpected solution to the adaptive problem. Two visual receptors were positioned geometrically such that visual fixation on the oblique edge of the triangle would typically result in a pair of visual signals (i.e., receptor 1 = low, receptor 2 = high) which was different from such pairs produced anywhere else in the arena (or rather *almost* anywhere else in the arena — see below). The robot would move in a straight line if the pair of visual signals was appropriate for fixation on the triangle, and in a rotational movement otherwise. Thus if the robot was fixated on the triangle, it would tend to move in a straight line towards it. Otherwise it would simply rotate until it did locate the triangle. Occasionally the robot would fixate, 'by mistake', on one edge of the rectangle, simply because, from certain angles, that edge would result in a qualitatively similar pair of visual signals being generated as would have been generated by the sloping edge of the triangle. Perturbed into straight line movement, the robot would begin to approach the rectangle. However, the looming rectangle would, unlike a looming triangle, produce a change in the relative values of the visual inputs (receptor 1 would be forced into a high state of activation), and the robot would be perturbed into a rotational movement. During this rotation, the robot would almost invariably refixate on the correct target, the triangle.

It is easy enough to see why Harvey et al.'s evolved triangle-seeking robot is a demonstration of non-trivial causal spread. In any orthodox cognitive-scientific model of triangle-square discrimination, the adaptive richness and flexibility of the intelligent behaviour would surely be

attributed to the systematic activity of neurally located representational states and computational processes. These might include perceptual inference rules, inner maps, route-calculating algorithms, and so on. In the evolved solution, by contrast, the observed richness and flexibility is secured by a system of organized interactions involving significant causal contributions not only from states and processes in the robot's nervous system, but also from certain additional non-neural bodily factors and from the environment. Thus one agent-side factor that deserves to be singled-out in the evolved solution is the spatial layout of part of the agent's physical body. This is not, of course, to deny neural events their rightful place in the agent-side causal story (even though the evolved neural network here was structurally quite simple). Rather it is to acknowledge the point that it is the geometric organization of the robot's visual morphology that is the primary factor in enabling the robot to become, and then to remain, locked onto the correct figure. The crucial part played by the environment becomes clear once one realizes that it is the specific ecological niche inhabited by the robot that enables the selected-for strategy to produce reliable triangle-rectangle discrimination. If, for example, non-triangle-related sloping edges were common in the robot's environment, then although the evolved strategy would presumably enable the robot to avoid rectangles, it would no longer enable it to move reliably towards *only* triangles. So the successful triangle-rectangle discrimination strategy depends not just on the work done by agent-side mechanisms, but also on the regular causal commerce between those mechanisms and certain specific structures in the environment which can be depended upon to be reliably present.

In contrast with its trivial cousin, non-trivial causal spread can be used to mount an anti-representational assault. Here's the argument:

1. In the present context, the distinctive target phenomena that inner representations are supposed to explain are the adaptive richness and flexibility of real-time intelligent action.
- So, 2. To qualify as action-oriented representations, neural states must account for that adaptive richness and flexibility.
3. Where real-time intelligent action is generated through non-trivial causal spread, factors in the non-neural body and the environment account for much of that adaptive richness and flexibility.
- So, 4. Any neural factors that causally contribute to the behaviour-generating process cannot count as action-oriented representations.

The extent to which one finds this argument compelling will probably be the extent to which one is in the grip of what Clark and I have previously called *strong instructionism* [43]. In the cognitive-scientific context, strong instructionism emerges as the view that what it means for some neural element to code for an item of intelligent behaviour is for that element to *fully specify* the rich and flexible aspects of that behaviour. If strong instructionism were the only candidate for an account of representation here, then the proposed sceptical argument would, it seems, be decisive, since the whole point about scenarios that feature non-trivial causal spread is precisely that some of the observed adaptive richness and flexibility turns out to be due to factors in the non-neural body and the environment. However, it really does seem that it ought to be rather easy to tame this particular anti-representational beast. We simply need to retreat from strong instructionism, and pursue an account of how neural states might make a distinctively representational contribution to on-line intelligent behaviour that does not impose the ‘full specification’ condition. In the next section I shall describe a strategy by which this

might be achieved. (Some alternative strategies, involving the properties of ‘being selected for’ and ‘decoupleability from immediate environmental input’, are discussed and rejected in [42].)

3 Representations: the Defence

I suggest that (i) being a genuine source of adaptive richness and flexibility, (ii) the presence of a certain sort of arbitrariness, and (iii) the existence of systemic homuncularity are jointly sufficient conditions for representation. And I suggest further that we can use this proposal to rescue representational explanation from the clutches of the causal-spread-wielding sceptic. This all stands in need of explanation.

As we have just seen, being a genuine source of adaptive richness and flexibility is certainly necessary for representation, because it is precisely the adaptive richness and flexibility of intelligent action which constitute the distinctive target phenomena that representations are ultimately supposed to explain. What this means is that if the non-trivial causal spread in some behaviour-producing system is such that *all* the observed adaptive richness and flexibility is traceable to extra-neural factors, then representational glosses of the neural contribution will be inappropriate. However, given what we know about the contribution of biological brains to intelligent action, such extreme situations will surely be few and far between. Nevertheless, being a genuine source of adaptive richness and flexibility is not sufficient for representation, because the whole point about cases of non-trivial causal spread is that causal factors located in the non-neural body and/or the environment may make that intelligence-related style of contribution. So if making such a contribution were sufficient for representation-hood, those extra-neural factors would count as representations, and that, I think, would be an unwanted

consequence, on grounds of excessive liberality.

Why should such liberality be deemed excessive? Previously (e.g. in [40]), I have suggested that it is enough here merely to emphasize the observation that the concept of representation has traditionally been used in cognitive science to help us explain the distinctive contribution of the brain, and to promote that observation into (what I have called) the *neural assumption*. The neural assumption is the constraint that however the concept of representation is ultimately to be understood, as a theoretical term in cognitive science, it must be restricted in its scope to neural states and processes. In fact, as an anonymous referee of this paper helped me to admit, the neural assumption is too strong. One thing that recent work in the area of situated cognition has brought into the foreground is the active causal role played by external representations (e.g. external linguistic resources) in generating and structuring human behaviour (for discussion, see e.g., [9, 10]). And part of the situated cognition message must be that the extra-neural factors highlighted here count as representational in a sense which is theoretically interesting to cognitive science. Nevertheless, there are limits on the kinds of extra-neural factors that ought reasonably to qualify as representations. At the very least, those qualifying factors should be (a) information-bearing resources and (b) play the role of standing in, within some behaviour-generating system, for some other entity or state of affairs (see above). But, of course, these styles of contribution do not exhaust the ways in which extra-neural factors might contribute non-trivially to real-time intelligent action. (Recall, for example, the specific way in which environmentally located shapes were seen to figure in an evolved solution to triangle-seeking.) I suppose we might conceivably be tempted here to drop the constraints provided by the functions of information-carrying and standing-in-for, and thus to open the floodgates to all manner of extra-neural entities counting as representations. But then we would be susceptible to an anti-

representational offensive based on parity considerations. This parity argument comes to the fore in the following reasoning due to Barbara Webb. In discussing her influential work on robotic models of mate-finding by female crickets, Webb says:

The robot operates without any attempt to build an internal model of its environment: there is no centralised representation of the sensory situation, not even in a distributed sense... It could be argued that the robot does contain 'representations', in the sense of variables that correspond to the strength of sensory inputs or signals for motor outputs. But does conceiving of these variables as 'representations of the external world' and thus the mechanism as 'manipulation of symbols' actually provide an explanatory function? It is not necessary to use this symbolic interpretation to explain how the system functions: the variables serve a mechanical function in connecting sensors to motors, a role epistemologically comparable to the function of the gears connecting the motors to the wheels. [37, p.53]

The pivotal thought here is that given parity of contribution, one must either treat the variables (neural factors) and the gears (extra-neural factors) as representational elements, or reject the language of representations altogether in favour of some less extravagant mechanistic terminology. Parsimony rules in favour of the latter option, as does our strong inclination to say that however important the contribution of the gears in connecting the motors to the wheels may be to adaptive success, that contribution is non-representational in character (cf. the contribution of the leg muscles in enabling animals to walk). (For more on the parity argument, see [41].)

To prevent any uninvited excessive liberality, we need to plug in our second and third conditions for representation, which identify two conceptually interlocking architectural features that

some behaviour-generating systems have and others don't, namely *arbitrariness* and *homuncularity*. Here I am using the term 'arbitrariness' in a very specific sense, according to which there is arbitrariness in a system just when the equivalence class of different inner elements that could perform a particular systemic function is fixed not by any non-informational physical properties of those elements (say their shape or weight), but rather by their capacity, when organized and exploited in the right ways, to carry specific items or bodies of information about the world (perhaps in an action-oriented form), and thereby to support an adaptive solution in which the information so carried helps to guide the overall behaviour [43]. The 'right ways' of being organized and exploited are established, I suggest, where the system in question is homuncular. And a system is homuncular just when it can be compartmentalized into a set of hierarchically organized, communicating subsystems, each of which performs a well-defined sub-task that contributes towards the collective achievement of an adaptive solution.

Considered separately, appeals to arbitrariness and homuncularity in this sort of context are not, of course, new. For the connection between arbitrariness and representation, see [32]. For the connection between homuncularity (or something very close to it) and representation, see, e.g., [3, 13, 20, 30, 36, 38]. However, the conceptually interlocking nature of these architectural features is not always appreciated. That interlocking nature becomes clear once one realizes that, in an homuncular analysis, the communicating subsystems are conceptualized as trafficking in the information that the inner vehicles carry. Indeed, it seems that the ways in which functionally integrated clusters of homuncular subsystems exploit inner elements, so as to collectively generate behavioural outcomes, are intelligible *only* if we acknowledge that those subsystems will have been set-up (in the natural world, by evolution or by learning) to have dealings with certain inner elements not because of those elements' non-informational

physical properties, but because of the information that they happen to carry. On this way of understanding things, certain subsystems are interpreted as producing information that is then consumed downstream by other subsystems.

Of course, homuncular subsystems must not be conceived as being, in any *literal* sense, understanders of the information in which they traffic. That would be to invite the debacle of an infinite regress of systems, each of which, in order to do what is being asked of it, must literally possess the very sorts of intentional capabilities (e.g., the capacity to understand the meanings of messages) that the model is ultimately supposed to explain. One traditional way of confronting this problem is to hierarchically decompose the overall system into arrangements of simpler and simpler subsystems, with a progressive simplification of function at each level of the hierarchy, until, finally, the sort of thing which you are asking each of your subsystems to do is something so primitive that the explanation is almost certainly going to be a matter for some kind of low-level neurobiology that doesn't appeal to information processing. It is this 'bottoming-out' in low-level neurobiology that is supposed to head off the threat of regress. As I have argued elsewhere [39, 42], this strategy, however much it is part of the received picture in the philosophy of cognitive science, is not beyond question. There remains, it seems, a mis-match between the brute physical causation apparently doing the discharging, and the interpretation-dependent causation which, as things stand, the homuncular metaphor seems to suggest operates at the higher levels of the hierarchy. Unless the transition between the two sorts of causation can be satisfactorily explained (or explained away), the homuncular metaphor seems to remain problematic. An alternative approach to the problem would be to develop an account of inter-subsystem communication which does not require that the information carried by the representations concerned is, in a literal sense, understood by those subsystems. To me,

this seems an achievable philosophical goal. (Here is not the place for the details, but one might point to the concept of communication that is deployed by animal behaviour theorists, when confronted by action-coordinating signalling transactions between certain non-human animals. For such creatures, it would be over-stretching our terms to say that the individuals concerned literally understand the meanings of the signals to which they respond. Nevertheless, a notion of communication still seems to have explanatory purchase (see, e.g., [15]).

What we have been describing is an economy within which certain elements *carry information* about external states of affairs (interpreted in an action-oriented way, as possibilities for action) in order to support behaviour-shaping communicative transactions between homuncular subsystems. Such an arrangement surely supports a description according to which the homuncular subsystems use the information-bearing elements to *stand in for* worldly states of affairs in their communicative dealings. And any such system must, it seems to me, warrant representational status. So I conclude that (i) being a genuine source of adaptive richness and flexibility, (ii) arbitrariness, and (iii) systemic homuncularity are jointly sufficient conditions for representation. Of course, the present proposal is secure only if the application of our three key conditions does not lead to the excessive liberality problem. In other words, we would have fallen short of our target of uncovering sufficient conditions for representation, if it were regularly true that, in systems featuring non-trivial causal spread, inappropriate component systems located in the non-neural body and the environment met the three conditions advocated. Such a turn of events cannot be ruled out *a priori*, but if one looks at the details of conditions (ii) and (iii), it really does seem an unlikely development. (For a worked-through example of how the proposed account of representation does locate some action-oriented representations, in the right place, in the midst of a behaviour-generating system rife with non-trivial causal spread,

see [40].) So far so good, then, for representation.

4 Snatching Defeat from the Jaws of Victory

I have argued that (i) being a genuine source of adaptive richness and flexibility, (ii) arbitrariness, and (iii) systemic homuncularity are jointly sufficient for representation. As it happens, I think they are necessary too. We have seen already that being a genuine source of adaptive richness and flexibility is necessary for representation. The additional necessity of arbitrariness is, perhaps, clear enough. Thus where the function in question is, say, holding my office door open, the equivalence class of suitable objects will be fixed by (roughly) the non-informational properties of being heavy enough and being sufficiently non-obstructive with respect to passing through the doorway. Here, where the equivalence class of different elements that could perform the function at issue is fixed by certain non-informational physical properties of those elements, there is simply no place for the language of ‘standing-in-for worldly states of affairs’ or, therefore, of representation. But now consider the function of keying my behaviour in rich and flexible ways to the door-stopping potential of some heavy book. The equivalence class of inner elements which may perform this function will be fixed precisely by the fact that certain inner elements are able, when organized and exploited in the right way, i.e., within an homuncular system, to carry some relevant item or body of information. Here it seems safe to say that the elements in question represent the associated worldly features. This suggests that arbitrariness is necessary for representation. And if, as I have suggested, arbitrariness and homuncularity arrive on the explanatory scene arm in arm (in that it is their interlocking nature which ensures that the target states are being used in the right sort of way to warrant a rep-

representational gloss), then the claim that homuncularity is necessary for representation looks to be concurrently established. (For a more detailed discussion of the necessity of homuncularity for representation, see [42].)

It is the fact that homuncularity is necessary for representation which opens the door to our second threat to representation. This new sceptical challenge can be put in the form of the following argument:

1. Homuncularity is necessary for representation.
 2. Modularity is necessary for homuncularity.
 3. Many of the biological systems underlying on-line intelligence are not modular in character.
- So, 4. So many of the biological systems underlying on-line intelligence are not homuncular in character.
- So, 5. So many of the biological systems underlying on-line intelligence are not representational in character.

As characterized above, a system is homuncular to the extent that it can be compartmentalized into a set of communicating subsystems in some hierarchical arrangement, and when each of those subsystems performs a well-defined sub-task that contributes towards the collective achievement of the overall adaptive solution. Given this account of homuncularity, it is a trivial observation that homuncular systems form a subset of *modular* systems, where, as I shall say, a system is modular to the extent that (a) it consists of scientifically identifiable subsystems, each of which performs a particular, well-defined sub-task, and (b) its global behaviour can be explained in terms of the collective behaviour of an organized ensemble of such subsystems. Given this account of modularity, homuncular systems are, straightforwardly, that subset of

modular systems in which the modules concerned (a) are hierarchically organised, and (b) can be said to communicate, rather than interact in some other way, with each other. So modularity is necessary for homuncularity. It follows that for the strategy of homuncular decomposition to be successful when applied to the control systems of intelligent agents, biological or otherwise, those brains must be seen to embody a particular sort of modularity, one that involves internal communications and is hierarchical in form.

As should be clear from the foregoing definition of modularity, I am here concerned with a *functional* rather than a *brutely physical* version of the phenomenon. This distinction can be readily illustrated if we think about biological brains. Such a brain will be functionally modular if it contains identifiable neural subsystems that perform specifiable sub-tasks. It will be modular in a brutally physical sense if it contains neural subsystems that can be picked out as having integrity because, for example, the interconnections between the neurons in the target group are dense while the connections to other groups of neurons are sparse. Of course, it may well be the case that functional modules in the biological brain (to the extent that they exist) will sometimes be realized by neural structures with anatomical integrity; but that is a further issue. This emphasis on identification by function also makes it clear that the neural structures which realize a module need not always be highly localized in space. Rather, those structures may be spatially distributed across regions of the brain. Nevertheless, it seems to me that there is, packed into the appropriate understanding of what it means here to pick out a subsystem that performs a well-defined sub-task, a requirement that we can assign a distinct functional contribution to some *part of* the overall system under investigation. In other words, we must be able to draw a spatial boundary around the very part of the overall system that, we contend, is the subsystem for performing the particular sub-task in question, even if that part happens to

be distributed in space. Call this requirement the *locatability condition*. Where it is not met, it seems to me that there may well be no empirical cash value to the claim that the system really performs (a) just those sub-tasks in just the way required by the proposed modular explanation, as opposed to (b) some alternative arrangement of sub-tasks identified by some other modular explanation that is behaviourally equivalent. (I take it that the plausibility of the locatability condition is, in part at least, what explains the theoretical ‘pull’ of brain-mapping techniques such as functional magnetic resonance imaging.)

Someone might respond here by complaining that spatial locatability is but one type of locatability, and that temporal locatability might do as well in enabling us to identify the kind of modules in which we are interested. As reasonable as this proposal sounds, I believe it to be misguided. Here’s why. Consider a system in which a single spatially identifiable region may perform a number of different specifiable sub-tasks — say tasks X, Y and Z. At any one particular time, however, this region may perform only one sub-task — X, Y or Z. How many modules for X do we have here? Using the temporal locatability condition alone, there seems to be no good reason to say (a) that we have just one module here, a module which performs task X on many different occasions, rather than (b) that we have an indefinite number of modules here, one for each time the region is activated in an X-related context. By contrast, plug in the spatial locatability condition and we get the surely preferable result (a), namely that we have just one module for X, a module which may be activated at an indefinite number of different times. Of course, this is not to say that spatial locatability is always straightforward. There may be fuzzy and/or dynamic boundaries in play. Nevertheless, the principle seems, to me, to be the right one.

If modularity is necessary for representation, it would be useful to have a generic character-

ization of the kind of causal system that will reward modular explanation. To do this, we can follow Clark [9, p.114] in appealing to Wimsatt's [44] notion of an *aggregate system*. As Wimsatt defines it, an aggregate system is a system in which (a) it is possible to identify the explanatory role played by any particular part of that system, without taking account of any of the other parts, and (b) interesting system-level behaviour is explicable in terms of the properties of a small number of subsystems. It seems clear that the essential qualities of aggregate systems make them ripe for modular explanation. The flip-side of the Wimsattian coin, however, is that non-aggregate systems will be resistant to modular explanation. Let's see how this works.

A system will become progressively less aggregative as the number, extent, and complexity of the interactions between its parts increases. Developing this thought, Clark [9] identifies the specific mode of causation that leads to such non-aggregativity as *continuous reciprocal causation*, defined as causation that involves multiple simultaneous interactions and complex dynamic feedback loops, such that (i) the causal contribution of each systemic component partially determines, and is partially determined by, the causal contributions of large numbers of other systemic components, and, moreover, (ii) those contributions may change radically over time. Using this terminology, we can state that the aggregativity of a system is negatively correlated with the degree of continuous reciprocal causation within that system. Now, as a system becomes less and less aggregative, with increasing continuous reciprocal causation, it will become progressively more difficult both to identify distinct and robust causal-functional roles played by reliably locatable parts of that system, and to explain interesting system-level behaviour in terms of the properties of a small number of subsystems, since the performance of any particular sub-task will increasingly be underpinned by larger and larger numbers of interacting components whose contributions are changing in highly context sensitive ways. Un-

der these circumstances, it is hard to see how the locatability requirement for modularity (see above) could be met. Indeed, as the sheer number and complexity of the causal interactions in operation increases, the grain at which useful explanations are to be found will become coarser. Thus our explanatory interest will be compulsorily shifted away from the parts of the system and their interrelations — and therefore, eventually, away from any modular analysis — and towards certain ‘higher-level’, more ‘holistic’ system dynamics (more on which later). So, the presence of continuous reciprocal causation undermines the practice of modular explanation. But now since modularity is necessary for representation, if some target system turns out to defy modular analysis, that system will be equally resistant to representational explanation.

All very interesting, you might be thinking, but is there any good empirical evidence that systems which perform useful tasks ever do feature continuous reciprocal causation? Time, once more, for evolutionary robotics to make its presence felt. For there is evidence that, *given certain kinds of evolutionary building block* (the primitives which evolution has to work with), Darwinian evolution will often produce adaptive control systems in which continuous reciprocal causation is harnessed to generate the adaptive solution. So what kind of evolutionary building block are we talking about? For some time, evolutionary roboticists have been keen to exploit the rich dynamical possibilities made available by so-called *dynamical neural networks* (DNNs). What we might, for convenience, call mark-one DNNs feature the following sorts of properties (although not every bona fide example of a mark-one DNN exhibits all the properties listed): asynchronous continuous-time processing, real-valued time delays on connections, non-uniform activation functions, deliberately introduced noise, and connectivity which is not only both directionally unrestricted and highly recurrent, but also not subject to symmetry constraints (see, e.g., [4, 6, 12, 14, 16, 21, 23, 27]). Recently, mark-two DNNs have added two further

twists to the architectural story. In these networks, which have been christened *GasNets* [26], the standard DNN model is augmented with (i) modulatory neurotransmission (according to which fundamental properties of neurons, such as their activation profiles, are transformed by arriving neurotransmitters), and (ii) models of neurotransmitters that diffuse virtually from their source in a cloud-like, rather than a point-to-point, manner, and thus affect entire volumes of processing structures (see, e.g., [18, 25, 26]). GasNets thus provide a mechanism by which evolution may explore potentially rich interactions between two interacting and intertwined dynamical mechanisms — virtual cousins of the electrical and chemical processes in real nervous systems. Diffusing ‘clouds of chemicals’ may change the intrinsic properties of the artificial neurons, thereby changing the patterns of ‘electrical’ activity, whilst ‘electrical’ activity may itself trigger ‘chemical’ activity.

So what happens when artificial evolution is given the task of designing GasNets which generate adaptive behaviour in robots? In a manner which is fully continuous with the selective regime described earlier, GasNet researchers typically allow artificial evolution to decide such fundamental architectural features as the number, directionality, and recurrency of connections, the number of internal units, and the parameters controlling modulation and virtual gas diffusion. In addition, certain aspects of the robots’ visual morphologies and certain motor parameters are usually placed under evolutionary control. It is striking that the underlying model of neural control present in GasNets seems almost to encourage evolution to harness the representation-resistant phenomenon of continuous reciprocal causation. Here we have DNNs which offer not merely the possibility of arbitrarily recurrent feedback loops and time-dependent dynamics, but also neurotransmitters that may transform the fundamental functional profiles (the transfer functions) of the neurons on which they act, and do so on a grand scale, given

that they act by gaseous diffusion through volumes of brain-space, rather than by electrical transmission along connecting neural ‘wires’. To be absolutely clear, the claim here is not that evolved GasNets will *always* realize adaptive solutions in which continuous reciprocal causation is harnessed (a claim that would fly in the face of the empirical evidence — see below), but rather that the fundamental processing architecture in play positively supports the possibility of such solutions.

That said, let’s consider what is at present the front-line empirical demonstration of GasNets at work, namely a series of re-runs of, and extensions to, the triangle-rectangle discrimination experiment described above [26]. Viewed as static wiring diagrams, many of the successful GasNet controllers appear to be rather simple structures. Typical networks feature a very small number of primitive visual receptors, connected to a tiny number of inner and motor neurons by just a few synaptic links. However, this apparent structural simplicity hides the fact that the dynamics of the networks are often highly complex, involving, as we would expect, subtle couplings between ‘chemical’ and ‘electrical’ processes. For example, it is common to find adaptive use being made of oscillatory dynamical sub-networks, some of whose properties (e.g., their periods) depend on spatial features of the modulation and diffusion processes, processes which are themselves determined by the changing levels of ‘electrical’ activity in the neurons within the network (for more details, see [26]). Preliminary analysis suggests that these complex interwoven dynamics will sometimes produce solutions which are resistant to modular decomposition. However, there is also evidence of a kind of *transient modularity* in which, over time, the effects of the gaseous diffusible modulators drive the network through different phases of modular and non-modular organization (Husbands, personal communication). It seems likely that the underlying causal story here is one in which the temporal unfolding of the system is characterized

by stages of relative aggregativity followed by stages in which continuous reciprocal causation takes hold.

For a second, and perhaps even clearer, demonstration of the fact that selection, when working with certain kinds of evolutionary building block, is likely to produce control systems that feature continuous reciprocal causation, let's move away from DNNs, and consider instead evolved control systems that feature a different kind of 'low-level' evolutionary primitive. For a number of years, Thompson [33, 34] has been pursuing a project in which artificial evolution is applied to reconfigurable electronic circuits. In motivating this research, Thompson highlights two constraints that are standardly placed on electronic circuits, constraints that are imposed with the aim of rendering those circuits amenable to human design. The first is now familiar to us: it is the "modularisation of the design into parts with simple, well defined interactions between them" [33, p.645]. The second is the inclusion of a clock: this gives the components of the system time to reach a steady state, before they affect other parts of the system. Thompson argues that once artificial evolution is brought into play, both of these constraints should be relaxed, since the richer intrinsic control-system dynamics that will result might well be exploited by evolution, even though human designers are unable to harness them.

For the roboticist to have any chance of exploiting the rich dynamical possibilities presented by abandoning the controlling clock, a problem must be overcome. That problem is that electronic components usually operate on time scales which are too short to be of much use to a robot. So one would like artificial evolution to be capable of producing a system in which, without there being any clock to control different time-scales, the overall behaviour of a whole network of components is much slower than the behaviour of the individual components involved. Thompson set artificial evolution this task, using, as his evolutionary raw material,

a population of (simulated) recurrent asynchronous networks of high speed logic gates. So although Thompson's networks do not actually generate intelligent robotic behaviour, they are certainly required to satisfy a condition that any unclocked system of that sort would have to satisfy in order to generate such behaviour. In this context, it is worth noting that biological nervous systems, like Thompson's recurrent asynchronous networks of high speed logic gates, feature no controlling clock of the relevant kind. Thus the way in which artificial evolution copes with what looks, from the perspective of human design, to be a challenging design problem may give us important clues to the way in which biological nervous systems work.

After forty generations (when the experiment was called to a halt even though fitness was still rising), a network had evolved which produced output that was over four thousand times slower than that produced by the best of the networks from the initial random population, and six orders of magnitude slower than the propagation delays of the individual nodes. What is striking is that the successful network seems to defy modular decomposition. As Thompson reports, the "entire network contributes to the behaviour, and meaningful sub-networks could not be identified" (p.648). Thus here we have a powerful example of artificial evolution producing a non-aggregate system, a system where (a) the complex nature of the causal interactions between the components means that "meaningful sub-networks [modules, functionally discrete subsystems] could not be identified", and (b) the system has to be understood in an holistic, non-modular manner ("the entire network contributes to the behaviour"). This is, of course, the distinctive stamp of modularity-threatening, homuncularity-threatening, representation-threatening continuous reciprocal causation.

5 The Search for Modularity

Representational explanation is no longer faring so well. But here is a rather obvious question: what will happen when evolutionary roboticists manage to evolve agents with behavioural capacities (or, perhaps even more to the point, suites of behavioural capacities) which are significantly more complex than those that have been evolved so far? Will they end up rediscovering control architectures that are usefully explained as being modular, homuncular, and representational? Resisting the temptation to engage in runaway speculations here, I shall offer a few brief and sketchy remarks on the issue of modularity, since even though not all modular explanations are homuncular explanations, modularity is necessary for homuncularity, and it is the modular decomposition of evolved control systems which is threatened directly by the presence of continuous reciprocal causation in such systems.

Everyone involved in evolutionary robotics agrees that long term progress in the field requires genetic encoding schemes that do not directly describe the entire network wiring in the genotype, but rather, as in nature, allow the building of multiple examples of a particular phenotypic structure, in a single ‘organism’, through a developmental process in which sections of the genotype are used repeatedly. Such developmental coding schemes are currently the target of feverish and exciting research (see, e.g., [7]). At first sight this might seem to wrap up the issue of control-system modularity in evolutionary robotics: the advent of richly modular genotypic encodings will lead to robust modularity at the phenotypic, control-system level, and then all that worrying about the lack of modularity in existing evolved DNNs, and in existing systems with evolved hardware, will turn out to have been misplaced. (This is, in my view, the most powerful way to raise a ‘scaling up’ worry for the anti-representational argument from continuous

reciprocal causation.) In fact, the issues here are far less clear-cut than this reasoning implies, since the question that matters for modular *explanation* is not ‘Will modular genetic encoding schemes produce repeated phenotypic structures in evolved control systems?’, to which the right answer is presumably a trivial ‘yes’. Rather, it is ‘Will such repeated phenotypic structures, when part of an advanced evolved control system, causally contribute to adaptive success in ways which (i) are not determined (to an overwhelming extent) by the causal contributions of large numbers of other structures in the system, and (ii) do not change radically over time?’, to which the right answer is, I think, ‘We’ll just have to wait and see’.

Of course, to be of scientific importance, continuous reciprocal causation need not be present in the target system *at all times*, or be present *throughout* that system. In the first place, as we have seen from the GasNet studies, a dynamically changing complex system may, over time, go through different phases of modular and non-modular organization. If, during the modular phases, the full conditions for representation are met, then that system will, at such times, be representational in character. By contrast, that same system will be non-representational in character during the non-modular phases. In the second place, the effects of continuous reciprocal causation may be restricted to certain regions of some complex system, so that we are presented with what is, at any one time, a hybrid organization of modular (potentially representational) and non-modular (non-representational) processes.

In this context it is crucial that even though the presence of continuous reciprocal causation may undermine the prospects for representational explanation, it does not necessarily put the offending system out of our explanatory reach. We just need a different sort of explanation. I do not have the space here to discuss this issue in detail (I say much more in [42]), but I can indicate where, I think, we need to look for an alternative explanatory framework. Of course,

if representation is necessary for computation (something which seems obviously true to me [42]), then computational explanation too will run aground here. But even this should not induce panic. Systems featuring continuous reciprocal causation surely look to be the kinds of systems that will often feature richly temporal phenomena such as adaptive oscillators (see, for example, the oscillatory dynamical sub-networks which, as we noted earlier, are often found in evolved GasNets). Such phenomena speak loudly in favour of the need to think in dynamical systems terms. Moreover, the presence of continuous reciprocal causation introduces pressure in the direction of a more ‘holistic’ explanatory perspective, since the componential interactions themselves will typically be too numerous and complex to capture in an illuminating or tractable way. Faced with this pressure, it seems plausible that one promising strategy may be to pursue an established style of dynamical systems analysis in which the variables that are identified as mattering for the explanation do not correspond straightforwardly to the properties of the internal components. Rather they constitute what have been called *collective* variables, in that they capture certain higher-order features of the target system. This allows the construction of a low-dimensional, higher-order state-spaces, which can then support explanations of the observed systemic behaviour (see, e.g., [5, 23, 28, 35]). It is, then, to dynamical systems theory that we may look in order to maintain our scientific (as opposed to merely our engineering) credentials.

Given that I have just promoted the use of collective variables and higher-order state spaces, an objection might be lodged against my claim that continuous reciprocal causation entails non-modularity. If we are permitted to think in terms of higher-level structures, what is to stop modules (and thus, given other factors, representations) turning up at that higher level of explanation, even in the case of systems that, looked at in a finer-grained way, exhibit all the

tell-tale signs of continuous reciprocal causation? Indeed, the objector speculates, such higher-level modules may even be revealed using the very style of dynamical systems approach that I have advocated. To see why this objection falls short, we need to recall the locatability condition for modularity. To genuinely pick out a module — a subsystem that performs a well-defined sub-task — we must be able to draw a boundary around (i.e., locate) some *part of* the overall system. As I argued above, the spatial route provides the most plausible way to identify, in a scientifically robust manner, something that could count as such a part (however distributed or temporally transient). But, according to the positive explanatory strategy which the proposed objection requires, it is possible for modular analysis to become radically disconnected from the details of the dynamic spatial organization of the target system — so far disconnected in fact, that the point that some systems, viewed in terms of *neural* function, do not reward modular thinking simply falls away as unimportant. Unfortunately, I do not see how this critic’s allegedly ‘modular’ explanation could ever satisfy the independently plausible locatability condition; so, I think, the objection fails.

6 Conclusions

Here’s what I think we can say with confidence about representational explanation, on the basis of our investigation into evolutionary robotics. As long as it turns out that neural resources are doing *some* of the intelligence-related work (and surely that is how the overwhelming majority of cases will turn out), non-trivial causal spread cannot be used to ground a robust anti-representational scepticism. A system involving non-trivial causal spread may still meet a set of conditions which are plausibly sufficient for that system to be representational in char-

acter. However, there is good evidence that when evolution is given the ‘right’ evolutionary building blocks (e.g., GasNets, reconfigurable electronic circuits), the result will often be control systems that feature high degrees of continuous reciprocal causation. Where that is so, those control systems will be stubbornly resistant to modular decomposition, and thus to homuncular decomposition, and thus to representational analysis. So now what about ALife-related robotics and representational explanation? Are they indeed friends reunited? It seems, from what we have seen, that evolutionary robotics remains a recalcitrant source of considerable disquiet. But then who said that a friendship can’t cope with the odd bit of friction.

Acknowledgments

All the examples of evolutionary robotics research which I discuss in detail in the text hail from the Evolutionary and Adaptive Systems Research Group at the University of Sussex. Many thanks, in particular, to Inman Harvey and Phil Husbands from that group for many hours of extensive discussion. I hope those hours were sometimes as useful for them as they were for me. Many thanks also to Mark Bedau and Andy Clark for making me think properly about the possibility of higher-level modules. I don’t expect the response I give here will satisfy either of them. Finally, this paper benefited greatly from the comments of two anonymous referees. Many thanks to them too.

References

- [1] Agre, P. E. & Chapman. D. (1979). What are plans for? In P. Maes, (Ed.), *Designing autonomous agents: Theory and practice from biology to engineering and back* (pp. 17–34).

Cambridge, MA, and London, England: MIT Press / Bradford Books.

- [2] Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48, 57–86.
- [3] Beer, R. D. (1995). Computational and dynamical languages for autonomous agents. In R. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp.121–47). Cambridge, MA: MIT Press / Bradford Books.
- [4] Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72, 173–215.
- [5] Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, 4(3), 91–99.
- [6] Beer, R. D. & Gallagher, J. G. (1992). Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1, 91–122.
- [7] Bongard, J. C. & Pfeifer, R. (2001). Repeated structure and dissociation of genotypic and phenotypic complexity in artificial ontogeny. In L. Spector & E. D. Goodman (Eds.), *Proceedings of the genetic and evolutionary computation conference, GECCO-2001* (pp. 829–36). San Francisco: Morgan Kaufmann.
- [8] Brooks, R. A. (1991). Intelligence without reason. In *Proceedings of the twelfth international joint conference on artificial intelligence* (pp. 569–95). San Mateo, California: Morgan Kauffman.
- [9] Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA, and London, England: MIT Press / Bradford Books.

- [10] Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford: Oxford University Press.
- [11] Clark, A. & Wheeler, M. (1998). Bringing representation back to life. In R. Pfeifer, B. Blumberg, J.-A. Meyer, & S.W. Wilson (Eds.), *From animals to animats 5: the fifth international conference on simulation of adaptive behavior* (pp.3-12). Cambridge, MA: MIT Press / Bradford Books.
- [12] Cliff, D., Harvey, I., & Husbands, P. (1993) Explorations in evolutionary robotics. *Adaptive Behavior*, 2, 73–110.
- [13] Dennett, D. C. (1978). *Brainstorms*. Brighton: Harvester Press.
- [14] Di Paolo, E.A. (2000). Homeostatic adaptation to inversion of the visual field and other sensorimotor disruptions. In J.-A. Meyer, A. Berthoz, D. Floreano, H. L. Roitblat, & S. W. Wilson (Eds.), *From animals to animats 6: Proceedings of the sixth international conference on simulation of adaptive behavior* (pp. 440–9). Cambridge, MA: MIT Press / Bradford Books.
- [15] Enquist, M. (1985). Communication during aggressive interactions with particular reference to variation in choice of behaviour. *Animal Behaviour*, 33, 1152–1161.
- [16] Floreano, D. & Mattiussi, C. (2001). Evolution of spiking neural controllers for autonomous vision-based robots. In T. Gomi (Ed.), *Evolutionary robotics: From intelligent robotics to artificial life*. Berlin and Heidelberg: Springer-Verlag.
- [17] Franceschini, N., Pichon, J. M., & Blanes, C. (1992). From insect vision to robot vision. *Philosophical transactions of the Royal Society, series B*, 337, 283–94.

- [18] Fujii, A., Ishiguro, A. Aoki, T., & Eggenberger. P. (2001). Evolving bipedal locomotion with a dynamically-rearranging neural network. In J. Kelemen & P. Sosik (Eds.), *Advances in artificial life: Proceedings of the sixth European conference on artificial life* (pp.509–18). Berlin and Heidelberg: Springer-Verlag.
- [19] Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.
- [20] Harvey, I. (1992). Untimed and misrepresented: Connectionism and the computer metaphor. Cognitive Science Research Paper 245, University of Sussex. Reprinted in *AISB Quarterly*, 96 (pp.20-7), 1996.
- [21] Harvey, I., Husbands, P., & Cliff, D. (1994). Seeing the light: Artificial evolution, real vision. In D. Cliff, P. Husbands, J.-A. Meyer, & S. W. Wilson (Eds.), *From animals to animats 3: Proceedings of the third international conference on simulation of adaptive behavior* (pp.392–401). Cambridge, MA: MIT Press / Bradford Books.
- [22] Haugeland, J. (1995/1998). Mind embodied and embedded. In his *Having thought: Essays in the metaphysics of mind* (pp. 207–37). Cambridge, MA, and London, England: Harvard University Press.
- [23] Husbands, P., Harvey, I., & and Cliff, D. (1995). Circle in the round: State space attractors for evolved sighted robots. *Robotics and Autonomous Systems*, 15, 83–106.
- [24] Husbands, P. & Meyer, J.-A. (Eds.) (1998). *Evolutionary robotics: Proceedings of the first European workshop, EvoRobot98*, volume 1468 of *Lecture notes in computer science*. Berlin: Springer-Verlag.

- [25] Husbands, P., Philippedes, A., Smith, T., & O'Shea, M. (2001). The shifting network: Volume signalling in real and robot nervous systems. In J. Kelemen & P. Sosik (Eds.), *Advances in artificial life: Proceedings of the sixth European conference on artificial life* (pp.23–36). Berlin and Heidelberg: Springer-Verlag.
- [26] Husbands, P., Smith, T., Jakobi, N., & O'Shea, M. (1998). Better living through chemistry: Evolving GasNets for robot control. *Connection Science*, *10*(3/4), 185–210.
- [27] Jakobi, N. (1998). *Minimal simulations for evolutionary robotics*. DPhil thesis, School of Cognitive and Computing Sciences, University of Sussex.
- [28] Scott Kelso, J. A. (1995). *Dynamic patterns*. Cambridge, MA, and London, England: MIT Press / Bradford Books.
- [29] Mataric, M. (1991). Navigating with a rat brain: a neurobiologically inspired model for robot spatial representation. In J.-A. Meyer and S. Wilson (Eds.) *From animals to animats: Proceedings of the first international conference on simulation of adaptive behavior* (pp. 169–75). Cambridge, MA: MIT Press / Bradford Books.
- [30] Millikan, R.G. (1995). *White queen psychology and other essays for Alice*. Cambridge, MA, and London, England: MIT Press / Bradford Books.
- [31] Nolfi, S. & Floreano, D. (2000). *Evolutionary robotics: the biology, intelligence, and technology of self-organizing machines*. Cambridge, MA: MIT Press.
- [32] Pylyshyn, Z. W. (1986). *Computation and cognition*. Cambridge, MA: MIT Press.
- [33] Thompson, A. (1995). Evolving electronic robot controllers that exploit hardware resources. In F. Moran, A. Moreno, J. J. Merelo, and P. Chacon (Eds.), *Advances in artificial life:*

- Proceedings of the third European conference on artificial life* (pp.641–57). Berlin and Heidelberg: Springer-Verlag.
- [34] Thompson, A. (1998). *Hardware evolution: Automatic design of electronic circuits in reconfigurable hardware by artificial evolution*. Berlin and Heidelberg: Springer-Verlag.
- [35] van Gelder, T. (1991). Connectionism and dynamical explanation. In *Proceedings of the thirteenth annual conference of the cognitive science society* (pp.499–503).
- [36] van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy*, *XCII*(7), 345–81.
- [37] Webb, B. (1994). Robotic experiments in cricket phonotaxis. In D. Cliff, P. Husbands, J.-A. Meyer, & S. W. Wilson (Eds.), *From animals to animats 3: Proceedings of the third international conference on simulation of adaptive behavior* (pp.45-54). Cambridge, MA: MIT Press / Bradford Books.
- [38] Wheeler, M. (1994). From activation to activity: Representation, computation, and the dynamics of neural network control systems. *Artificial Intelligence and Simulation of Behaviour Quarterly*, *87*, 36–42.
- [39] Wheeler, M. (1998). Explaining the evolved: Homunculi, modules, and internal representation. In P. Husbands and J.-A. Meyer (Eds.), *Evolutionary robotics: Proceedings of the first European workshop, EvoRobot98*, volume 1468 of *Lecture notes in computer science* (pp.87-107). Berlin: Springer-Verlag.
- [40] Wheeler, M. (2001). Two threats to representation. *Synthese*, *129*, 211–31.

- [41] Wheeler, M. (Forthcoming). How to do things with (and without) representations. In R. Menary (Ed.), *The extended mind*. John Benjamins.
- [42] Wheeler, M. (Forthcoming). *Reconstructing the cognitive world: the next step*. MIT Press.
- [43] Wheeler, M. & Clark, A. (1999). Genic representation: Reconciling content and causal complexity. *British Journal for the Philosophy of Science*, 50:1, 103–35.
- [44] Wimsatt, W. (1986). Forms of aggregativity. In A. Donagan, N. Perovich, & M. Wedin (Eds.), *Human nature and natural knowledge* (pp.259–93). Dordrecht: Reidel.