Towards a Cognitive Robotics

ANDY CLARK & RICK GRUSH
Washington University

There is a definite challenge in the air regarding the pivotal notion of internal representation. This challenge is explicit in, e.g., van Gelder, 1995; Beer, 1995; Thelen & Smith, 1994; Wheeler, 1994; and elsewhere. We think it is a challenge that can be met and that (importantly) can be met by arguing from *within* a general framework that accepts many of the basic premises of the work (in new robotics and in dynamical systems theory) that motivates such scepticism in the first place. Our strategy will be as follows. We begin (Section 1) by offering an account (an example and something close to a definition) of what we shall term Minimal Robust Representationalism (MRR). Sections 2 & 3 address some likely worries and questions about this notion. We end (Section 4) by making explicit the conditions under which, on our account, a science (e.g., robotics) may claim to be addressing cognitive phenomena.

keywords: representation, forward models, anti-representationalism, reactive systems, motor control, feedback.

0. INTELLIGENCE & REPRESENTATION

Contemporary cognitive science, it is fair to say, displays a deep-seated commitment to a representational view of the mind. According to such a view, intelligence is largely a matter of problem solving, and problem-solving is carried out via computations defined over internal representations of salient real-world structures, facts and hypotheses. Cognitive science then earns its status as a special science of the mind by devoting itself to the distinctive projects of tracing the transformations and flow of these internal encodings and of understanding the processes that mediate such transitions.

This picture may be dubbed, without malice, the same old story (SOS). Classic statements of SOS include, e.g., Pylyshyn, 1987; Fodor 1975, 1987. But the same broad outline applies equally to the bulk of work in connectionism and neural networks (Rumelhart, McClelland, & The PDP Research Group, 1986; Smolensky, 1988; Elman, 1993; Churchland & Sejnowski, 1992). Nevertheless, scepticism concerning SOS is undoubtedly on the rise. In particular, there is a definite challenge in the air regarding the pivotal notion of internal representation itself. Very roughly, the challenge is to show:

1. That the operative conception of internal representation is non-trivial.
2. That identifying certain internal states as representations does explanatory work.
3. That such identification is empirically possible.
4. That the kinds of states so identified figure deeply in biological cognition—that they are not just some tip of the iceberg phenomenon.

This challenge is explicit in, e.g., van Gelder, 1995; Beer, 1995; Thelen & Smith, 1994; Wheeler, 1994; and elsewhere. We think it is a challenge that can be met and that (importantly) it can be met by arguing from *within* a general framework that accepts many of the basic premises of the work (in new robotics and in dynamical systems theory) that motivates such scepticism in the first place. Our strategy will be as follows. We begin (Section 1) by offering an account (an example and some-

AC: Philosophy/Neuroscience/Psychology Program, Dept of Philosophy, Washington University, St. Louis, MO 63130, USA. andy@twinearth.wustl.edu

Copyright 1999 International Society for Adaptive Behavior

thing close to a definition) of what we shall term Minimal Robust Representationalism (MRR). Sections 2 & 3 address some likely worries and questions about this notion. We end (Section 4) by making explicit the conditions under which, on our account, a science (e.g., robotics) may claim to be addressing cognitive phenomena.

1. MINIMAL ROBUST REPRESENTATIONALISM[1] (MRR)

Consider the class of real-world, real-time actions known as skilled reaching (also known as *fast voluntary goal-directed movement*). Skilled reaching is the smooth approach of an arm and hand system towards some target object.[2] Success in this class of actions depends, in part, upon the brain's receiving and responding to a stream of proprioceptive feedback, especially when visual feedback is not available: feedback concerning the orientation, position and trajectory of the arm/hand system as the movement progresses. There is, however, a widely appreciated and seemingly insurmountable problem. The proprioceptive feedback is often (for very fast movements) required, it seems, faster than it is available. For such feedback to be used to smooth-out fast on-going reaching activity, it needs to be available before the minimum naturally possible delay has elapsed.[3] Yet smooth reaching happens! How does nature turn the trick?

What we confront is, in fact, an instance of a quite general problem faced by many systems involved in the delicate, real-time control of distal processes. For many controllers, it seems, are required to make delicate adjustments based on information that is just not available fast enough to be used to modulate the process. This occurs in, for example, chemical plants that need to control an on-going reaction by adding chemicals to a mix but where waiting for feedback cues to prompt the process is impossible since, by the time the cues are received, it would be too late for the chemical infusions to work. The same situation arises in bio-reactors that must keep the bio-mass in a tank constant by delicately altering nutrient inflow, heat and stirring rate (Ungar, 1990).

One solution, sometimes used in industrial settings such as those just described, is to use an *emulator* to provide a kind of quicker mock feedback for use by the control system.[4] An emulator is just a mechanism (circuitry, software routine, whatever) that takes as *input* information about the starting (or current) state of a system (e.g., biomass, temperature, etc.) and about the control commands that are being issued (e.g., increase heat by 2 degrees). The emulator then gives as *output* a prediction of the next state of the system. This prediction takes the form of a set of values for the future feedback that the new state of the system should yield. The emulator thus *models* the target system and generates a kind of mock feedback that can be used instead of the laggardly feedback from the real system. In the reaching scenario described earlier, this would involve the provision of additional on-board neural circuitry that takes a copy of various motor commands as input and is set up (by learning or design) so as to replicate the relevant dynamics of the real bodily system. The output of the emulator is thus, in this case, a quick prediction of the future proprioceptive feedback from the hand/arm system (Ito, 1984; Kawato et al., 1987; Wolpert et al., 1995). How is such a prediction achieved?

Consider the real arm during a typical movement. It exhibits a range of dynamical behaviors (behaviors that change over time) including changes to shoulder and elbow angle, changes in the angular velocity and acceleration of the elbow joint, changes in the angular inertia of the arm, changes in the agonist and antagonists torques to the shoulder joint and so on. An emulator circuit for an arm control system might thus consist (this is the maximally simple case—see later for complications!) in a simple neural network comprising a number of connectionist-style units. Each individual unit would be devoted to a different one of the parameters just discussed and the inter-unit connectivity set up so as to replicate the interactions and interdependencies of the various parameters (e.g., angular inertia rises as elbow angle decreases). This simple vision is not overly fanciful. Kawato (1990) trained a neural network to plan arm trajectories and found, on subsequent analysis, several units whose response profiles tracked the evolution of very specific physical parameters such as those mentioned above. The upshot is thus a mini dynamical system in which one unit stands for each parameter, each unit is able to reproduce the evolution equation of its associated parameter, and the inter-unit connectivity mirrors the physical interdependence between the parameters.

The provision of such emulators[5] can enhance the functionality of a system in a wide variety of ways. It allows the system to exploit mock feedback signals available ahead of the real-world feedback, and hence allows rapid error-correction and control. It can support reasonably sensible behavior in the total absence of real-world feedback. And it can allow the improvement of motor skills to continue off-line - the agent can practice using the emulator model without engaging the real-world system at all. All these benefits seem visible in the human case. And damage to the real neural circuitry plausibly associated with such emulator functions causes reaching to become shaky and oscillatory (intention tremor), as you would expect if the system is then reduced to relying on slow, real-world proprioceptive feedback. Moreover, as just indicated, the neural basis of motor emulation is beginning to be understood. It looks to implicate circuitry involving the cortico-spinal tract, the red nucleus, the inferior olive and the contralateral dentate and cerebellar cortex (Kawato, Furukawa & Suzuki, 1987; Dean, Mayhew & Langdon, 1994). In this regard it is interesting to note that neurons in the magnocellular red nucleus of the cat have been found to mimic the dynamics of specific parameters of cat forelimb motion (Ghez, 1990; Ghez & Vicario, 1978).

Time, then, to cut to the chase. Our case for a minimal robust representationalism begins with the following claim: that analytically traceable instances of emulator circuitry would constitute the most evolutionary basic scenario in which it becomes useful to think of inner states as full-blooded representations of extra-neural (in this case, bodily) states of affairs. Just what we mean by a full-blooded representation will become clearer later on. For now, notice only that an emulator circuit such as the toy neural model described earlier would surely provide:

1. A pretty clear case of a system of inner states and processes whose adaptive functional role is to *stand in* for specific extra-neural states of affairs.

2. A type of system that allows the *precise identification* of component states and processes with representational roles (the units each model a specific physical parameter, and the connection weights track the physical interdependencies between parameters).

3. A compelling demonstration of a case in which the provision of an inner model is not an impediment to real-time success but actually *enhances* fluent real-time action.

This last point is aimed at an earlier and influential argument due to Rodney Brooks (Brooks, 1994)) to the effect that inner models are a costly information-processing bottleneck and an impediment to fluent, real-time success. The motor emulation case may thus act as a kind of bridge between the real-world, real-time, on-line focus of recent work in robotics and autonomous agent theory (Brooks, 1994; Beer, 1995; Agre & Rosenschein, 1996) and the more traditional Cognitive Scientific focus on inner models and de-coupled reasoning. For it speaks to the question, how did the process of genuinely representing the world (as opposed to simply interacting with it) ever get started? How did internal representation (in a strong sense, still to be defined) ever get its foot in the door of real-world, real-time cognition? The answer, we speculate, is that world-modeling got its foothold when nature discovered that emulator circuitry could improve real-world, real-time responsiveness. With that circuitry on hand (so to speak) it probably required only minor cheap modifications to glean the added benefits available from running the emulator completely off-line so as to aid planning, support mental imagery, and so on. Early emulating agents would then constitute the most minimal case of what Dennett calls a Popperian creature — a creature capable of some degree of off-line reasoning and hence able (in Karl Popper's memorable phrase) to "let its hypotheses die in its stead" (Dennett , 1995, p. 375). It would also, by the by, constitute a phylogenetically basic case of what Annette Karmiloff-Smith has dubbed representational re-description —the creation of a new, more manipulable inner space on top of some basic sensorimotor routine (Karmiloff-Smith, 1992; Clark & Karmiloff-Smith, 1993). In addition, emulators seem to be a nice, biologically detailed example of the sort of disengagement that Brian Cantwell Smith (1996) has recently argued to be crucial for understanding representation.

In presenting emulation as a minimal yet robust case of internal representation, we are, we believe, building on a common and intuitive sense of what representation should involve. Consider central and familiar cases of external representations, such as Wittgenstein's example of the use of model cars in a courtroom re-construction, or the town planner's use of models and mock-ups to aid problem-solving and facilitate debate. In all these cases, the external structures are properly said to *represent* absent, distal or counterfactual states of the real-world because they are manipulable structures whose functional role is to vary and interact in ways that directly correspond to actual or potential variations and interactions in the real-world arenas of interest.

Our suggestion is that this is a useful place to situate one fuzzy *but important* boundary that distinguishes various forms of adaptive agency (Campbell, 1974). For some adaptive systems, then, use models (internal and external) in place of directly operating upon the world. Non-cognizers, by contrast, remain trapped in a (potentially very complex and context-variable) web of closed-loop interactions with the very aspects of reality upon which their survival depends.

One virtue of this way of drawing the distinction is that *what* gets internally represented and *how* it gets represented are left very much up for grabs. For example, in our toy case, it is noteworthy that the representations are rather closely tied to the real bio-mechanics of the system, that they are egocentric (and indexical in the sense of Agre (1996)), and that they can be effectively implemented using connectionist-style resources and analog components (representation and digitality thus come apart, pace e.g., Stufflebeam (1996)). The toy system thus occupies an interesting region of logical space insofar as it is quite distant from the GOFAI (Good Old Fashioned Artificial Intelligence—(Haugeland, 1985)) vision of central codes, language-like representations and so on, yet still looks to present a compelling instance of a representation-using system.

In sum, our suggestion is that a creature uses full-blooded internal representations if and only if it is possible to identify within the system specific states and/or processes whose functional role is to act as *de-coupleable surrogates* for specifiable (usually extra-neural) states of affairs. Motor emulation circuitry, we think, provides a clear, minimal and evolutionarily plausible case in which these conditions may be met. In addition, emulator circuitry helps explain the complex patterns of connection and disconnection between agents and the objects of their thoughts that, according to Brian Cantwell Smith (1996), are crucial for understanding the nature of representation itself. It remains, however, to further clarify the implicit contrast between full-blooded and weaker notions of internal representation and to respond to some likely worries about our approach. This de-bugging exercise is the topic of the next section, to be conducted in a kind of 'question-and -answer' format.

2. DE-BUGGING MINIMAL ROBUST REPRESENTATIONALISM

Comment: "What the emulator circuit does is very neat—but that's not what I mean by internal representation."

Response: We agree that it certainly isn't classical, GOFAI-style representation. But given the ease with which we can associate quite specific inner activities in the circuitry with the functional role of standing in for specific extra-neural states of affairs, it seems perverse to deny that the inner states are acting as representations. It is, of course, true that the emulator circuitry can *also and simultaneously* be viewed simply as a smaller dynamical system linked to the one that hooks directly into the real-world. But that's just as it should be. The circuitry can also and simultaneously, be viewed as a mass of quarks and electrons or as a quantum wave packet. The question is, which of these descriptions is most useful for Cognitive Science?

Comment: But that's just the point. The representational descriptions *aren't* helpful—they don't add anything to the understanding you would get by knowing the dynamics alone.

Response: This is a tricky one, since folks do vary in what gives them that warm glow of explanatory satisfaction. But let's try. We suggest that the representational glosses reveal something about the relation between the system's structure and the tasks it is supposed to perform. If we ask why a given unit has the response profile it does, or why unit x is connected to unit y in just the way it is, then learning that the units are standing in for the real-world evolution of two specific arm motion parameters is surely illuminating. Representation-talk, in this case, is the glue that binds *telos* (purpose) to mechanism. Importantly, this binding is achieved in a way that simultaneously helps explain the physical characteristics of the specific target device (the actual emulator circuit) and that displays—in a necessarily abstract way—something of the space of other possible devices that could fulfill the same role. Thus the representational description helps fix the equivalence class of dynamical systems that could fill the specific functional role of enabling smooth real-time reaching in the absence of sufficiency fast proprioceptive feedback. This equivalence class will include circuits that look very different when viewed without the lens of our representational understanding and will include e.g., circuits that use multiple units to code for each parameter, circuits that use over-lapping populations of units and 'coarse coding', ones that use simple look-up tables, and ones that exploit more complex dynamical features such as limit cycles. Considered just as dynamical systems, these are all very different critters indeed. A representational gloss, we suggest, helps us isolate the functionally salient properties in virtue of which all these circuits are fit to play the same abstract role in a wider system: a role defined by reference to what the circuitry is for and what it helps to achieve. Surely that is why it is exciting to learn e.g., that certain neurons in the cat's magnocellular red nucleus mimic the dynamics of forelimb motion—it is exciting because it is a clue (though only a clue) to what the neurons are really doing for the system.[6] Such clues are of great practical importance as they influence the way we go on to conduct future research e.g., we look at other circuits quite closely connected to this one and ask what they might be doing that draws on this kind of information and so on.

Comment: Even if there is some heuristic value (to *us*) in analyzing the system in representational terms, we should not be misled into believing that those units (or whatever) actually represent anything to the system.

Response: This, as it happens, is exactly the kind of worry that our account is designed to meet. It s a worry that has recently surfaced in a number of places. Thus consider the following comments:

> " . . . if a particular neuron fires . . . whenever

> something red is visible, that neuron. . . may be said to represent the presence of something red. While this argument may be perfectly reasonable as an observer's explanation of the system, it should not be mistaken for an explanation of what the agent in question believes." (Brooks & Stein,1993, p. 2)

> "Computer models of autonomous agents contain symbolic structures that represent theoretical entities to the modeler, while computational theories of autonomous agents claim that agents contain symbolic structures that represent their situation to themselves." (Beer,1995, p. 175)

At least one of the present writers (Clark) used to dismiss this kind of comment as trading on the mistaken idea that every representation (to be a representation for the system) needs to be conceptualized *as* a representation *by* the system: a feat which clearly calls for very sophisticated reflection and looks likely to bring in considerations of conscious awareness, linguistic abilities and so on. Alternatively, it might be thought that all that is being pointed out is that we should not be too quick to infer function from correlation. The mere fact that some inner state correlates with the presence of red (or whatever) falls well short of establishing that its systemic role is that of red detector —see e.g., Churchland & Sejnowski's (1992, pp. 185-186) lovely discussion of units that correlate with the presence of edges but whose systemic role is to extract curvature from shaded images. Let us assume, however, that we are not misled about systemic role, and that the critic is not merely drawing attention to the fact that the system is unaware of its own representational activity. In such a case, is there any further value in distinguishing between representation-to-us and representation-to-the-system?

We think that there is, and that what is really at issue is an important and illuminating contrast between what we will term (boringly, but descriptively) weak versus strong senses of internal representation. Thus consider an arm control system that must operate without the benefit of additional emulator circuitry. The system uses a mixture of ballistic early motion and whatever error corrections can be performed using only the slow, proprioceptive feedback available from the real arm and hand system itself. Such a system might well involve a number of inner states that can be correlated with real-world events. And these correlations are not accidental—it is indeed the adaptive role of these inner states to thus track and respond to changes in arm orientation. Nonetheless none of these states function as *stand-ins* for e.g. arm angle in the strong sense of playing the part of arm angle in such a way that the real-world arm movements need not *themselves* participate in various episodes of control and coordination. The difference is thus between inner states that continuously link the processing to the on-going evolution of extra-neural reality and inner states that recapitulate the dynamics of extra-neural reality without depending on a constant physical linkage between the inner states and what they are about. Only the latter constitute cases of what we are calling 'strong internal representation.' The former are states whose role is not to model the world as such (that is the province of strong representation) but rather to guide our activities by creating closed loop interactions that yield adaptively valuable responses keyed to the on-going evolution of the physical world as it is impacting upon our sensory peripheries.

The problematic distinction (between systems that can be usefully understood, from the outside, in representational terms and those in which the representations are actually for the system) is thus revealed as genuine. For systems such as the simple, non-emulator based arm controller may successfully deal with some domains without needing to create independently manipulable items capable of standing-in for elements of the domain outside the loop of the system's direct interactions with its world. It is this distinction, we submit, that has led the more sensitive philosophers of cognitive science (Haugeland, 1991; Smith, 1996) to tie the idea of internal representations to ideas involving tracking the distal, the absent, and the non-existent. For such cases look to require the use of stand-ins in our strong sense i.e., inner items capable of playing their roles in the absence of the on-going perceptual inputs that ordinarily allow us to key our actions to our world.

The upshot is that ordinary controllers (that is, controllers which do not employ emulator circuitry) do not, in our strong sense, represent the events they control—no more than, to take a mundane example, the sculptor's hand represents the clay it manipulates. We are thus led to re-invent the distinction between real representational systems and what Israel (1988) called 'information and control systems.' These are any systems in which transduced information plays the adaptive role of promoting successful goal-oriented behavior. Such systems may indeed be apt for external analysis in terms that describe such and such a neural population as representing such and such a state of affairs. But unless the system displays, in addition, the capacity to set-up and manipulate inner models instead of operating directly upon the world, it will fail to count as a locus of full-blooded internal representation. Chess computers that 'try out' possible moves so as to assign values to options and then act on the basis of such comparisons are thus full-blooded representers,[7] whereas a simple phototropic (light following) robot is not. Similarly, the tic-tac-toe network (McClelland et al., 1986) that uses it's own play module to simulate an opponent's possible responses before choosing a move involves a case of full-blooded representation, whereas a tic-tac-toe network with a hard-wired response for every possible move does not.

Comment: The more I think about the real motor emulation case, the less clear I am on its role in the argument. Such emulator circuitry seems to be working just one step ahead of the real-world feedback. This falls well short of an inner model that can be consulted fully independently of the system's ongoing interactions with the world.

Response: This worry was raised, in somewhat different forms, by two anonymous referees. We think it is a good one, as it allows us to clarify our view of the relation between full-blooded internal representation, minimal robust representation and the weasel word 'cognition'. In Section 1 we now identify full-blooded internal representations with *fully de-coupleable* inner surrogates for extra-neural states of affairs. The case of basic motor emulation does indeed fall short of meeting this stricter criterion, and for just the reasons mentioned in the comment: the surrogate states are not fully-decoupleable from ongoing environmental input. Instead, the surrogates act to provide a kind of fine tuning for environmentally coupled action. Such cases constitute the most minimal example of a representational strategy. In fact, we do not mind *how* these minimal states are described. What matters is that the basic strategy of using inner states to stand in for (in this case temporarily) absent states of affairs is here in place. We have arrived at the most minimal end of what is surely a rich continuum of possible 'stand-in invoking' strategies.

Our claim is that the kind of circuitry that supports this basic functionality provides the basis upon which stronger forms of internal representation can then develop (hence our speculations concerning the fully off-line use of the same emulator loop). An ability to deploy the same circuitry without any actual activity (perhaps by simply issuing an action command and actively inhibiting bodily response) then moves us up the representational continuum. It allows the emulator-exploiting system to predict the shape of possible actions much further ahead (not just one step into the future). At the top end of this continuum lie creatures capable of manipulating rich inner models of many possible environments and actions and capable of using those manipulations to plan ever- more effective and complex courses of action.

This same continuum-based view applies to the vexed question of the cognitive/non- cognitive divide. We think it is misleading to insist on a neat divide. But there are genuine differences among adaptive strategies. And the strategies that allow creatures to best cope with the distal and the absent are, traditionally, those most strongly associated with the notion of a *cognizant* agent. Now it is true (as one referee usefully points out) that not even the sceptic about internal representation means to deny the obvious fact that human agents can and do think about the distal and the absent, that we can and do make plans involving future chains of events etc. So the mere presence of such capacities cannot be the object of our theoretical discussion. Instead, we are suggesting that, as a matter of empirical fact, the capacities most strongly associated with the traditional notion of cognition will turn out to be supported (made possible) by the presence of robustly representational inner states or processes: inner states or processes that are a) scientifically (non-semantically) identifiable and b) serve as stand-ins for specific extra-neural states of affairs. If no such identifiable substates or processes are to be found (perhaps because the system is too unstable, too complex, or too holistic in operation) then we would have to concede the anti-representationalist point. And we do, in any case, anticipate a kind of continuum effect.

Some adaptive strategies will rely on inner states that are more *strongly* representational than others (e.g. full and free off-line useability of the stand-ins, versus cases where the use is more constrained, either to a context or to a sequence or to a certain depth of look-ahead). And some, as we saw, will not involve the use of stand-ins at all. Our suggestion is that the cognitive/non-cognitive divide will turn out to be empirically tied to the presence or absence of identifiable systemic stand-ins. The move 'towards a cognitive robotics' will thus be marked by an increasing reliance on strategies such as off-line modeling and emulation.

Does it follow (as one referee suggested) that it now becomes an empirical question whether *we humans* are indeed cognitive agents? This would indeed be odd, though it might become more palatable *once* it is accepted that the scientific issue concerns not so much *what* we can do (e.g. think about the distal and absent) as *how* we can do it. But we need not go so far. Instead, our claim is that *as a matter of fact* the capacity to think about the distal and the absent is grounded in the use of systemic stand-ins and emulation-based strategies. In other words, there *could* (logically) be non-(strong) representation using cognizers.[8] But biological systems are just not like that.

It is also worth noting, however, that the use of such strategies and stand-ins is meant as a necessary and not a sufficient condition for the attribution of a cognitive competence. So a simple robot that uses off-line world models may still fail to qualify as a cognitive agent. (There may, for example, be additional constraints involving flexibility and environmental complexity.)

Comment: But isn't this a rather revisionist take on internal representation? — Popular cognitive scientific usage seems so much more lax.

Response: There can be no doubt but that the neuroscientific community, at least, tends to talk of internal representation when all that is really at issue are states of 'information and control.' By contrast, philosophers (and cognitive scientists concerned with higher-level reasoning, planning, story-understanding and so on) have often been guided by something much more akin to our notion (Haugeland, 1991; Smith, 1996). We now think that some degree of revisionism is inevitable, and that one of the more positive lessons of recent debates about internal representation (such as the exchange between Vera & Simon and their critics (Vera & Simon, 1993; Touretsky & Pomerleau, 1994) is that *something really does have to give*. For the cognitive scientific usage of the term representation is trying to do too much. It is torn between 1) a notion that is tied to the idea of inner states that merit external understanding as information-bearers (an account that covers sun-tracking plants, low level vision and much else besides (Smith, 1996), and 2) a notion tied to the idea of inner states whose character helps us to distinguish the real cognizers from the rest—a distinction that, we claim, involves identifying systems capable of breaking their cycle of direct interactions with the world and reasoning outside the loop . We don't see how (except by oscillation, ambiguity or self-deception) the term 'internal representation' can figure in both these projects at once. In opting for the stronger, model-based reading we must perforce embrace a degree of revisionism. What we get to keep is the idea of representations as inner surrogates that allow some animals to engage in the vicarious (hence safer) exploration of their worlds (Campbell, 1974). What we must give up is the idea that all cases of feature detection, coordinate transformation, etc. necessarily involve internal representation *in the same sense*.

Comment: Even if the (strong) representational story helps us to understand what is going on in the emulator circuit, it is surely not a sufficient explanation. We must also attend to the real temporal dynamics that allow the circuit to participate in the real-time control of motion.

Response: We totally agree. We must indeed face up to the fact that the effectively atemporal Turing machine model of computation leaves out much that is of interest for understanding biological cognition. It is clear that, for example, the adaptive value of the emulator circuit depends as heavily on the precise real-time profile of its response as upon the representational roles of the units. A satisfying analysis of the emulator circuit will thus involve both an understanding of the representational roles of its states and processes and an account of the real-time circuit dynamics that fit the temporal profile of the emulator to the temporally evolving needs of the overall system. The detailed understanding of some strongly representational systems may thus involve flipping back and forth between a traditional representational approach and something more like a dynamical systems analysis (Port & van Gelder, 1995).

Comment: Can't any reasonably sophisticated system be described as an emulator? And if so, then doesn't this render the project of finding biological emulators a pointless exercise?

Response: These comments, made by another anonymous referee, allow us one final chance to clarify some crucial points. Actually, the requirements for being an emulator are rather strict. First, there must be a control loop of some sort, such that one part of the loop is usefully identified as the controller, and the rest identified as the controlled system. Examples of such control loops are thermostat/room systems, car/driver systems, and even brain/body systems. Then, there must be some additional system which can accept a copy of the control signal produced by the controller and on that basis produce a mock version of the feedback signal that the controlled system would produce if it were to process that same control signal. These requirements are far from trivial. Not everything is a flight simulator, exactly because not everything is so constructed so as to take typical pilot control signals (joystick movements, etc.) and produce mock versions of the feedback that a real aircraft would provide (instrument readings, visual scene, etc.).

As far as the identification of biological emulators goes, presumably one would look for neural pathways which apparently carry efferent copy information. These would be neural lines which break off from normal efferent lines, and carry copies of the efferent signal elsewhere in the central nervous system for processing. One would then see to where these efferent copy lines lead, and see if there is any evidence that what they are doing is processing these copies to produce mock feedback. Does damage to these areas compromise fast voluntary movements (an appropriate question for the sort of motor emulator discussed in Section 1), or more generally, does damage to that area produce the sorts of deficits one would expect if that area is acting as a neural stand in of this or that sort? These questions may not be easy to answer, and there will often be alternate interpretations of even the most seemingly straight-forward data (see, e.g., Footnote 6). But this is how it is with all attempts to identify the function of neural systems. We address this issue a bit more in the next section.

3. BURGEONING COMPLEXITY?

Our claim, in a nutshell, is that there is a strong (non-trivial) sense of internal representation that is both explanatorily potent and that looks to be quite widely applicable to real biological systems—the sense, namely, of inner states and processes whose functional role really does involve playing the part of environmental features that need not be in close, constant interaction with the agent at the time. This notion answers to all the demands laid out in our opening section save one: the demand (number 3 on our original list) that we be able to identify in biological organisms the specific inner states or processes that play particular representational roles. Is such identification empirically plausible once we leave the realm of toy devices and confront biological systems in all their gory complexity? In short we now confront the most worrying skeptical response of all: okay, as far as the emulator circuits go, but *real brains just aren't like that.*

This is a genuine worry and one for which there is, as yet, no compelling remedy: only time and advances in neuroscience will tell. Certainly we accept the bare possibility that biological brains might constitute systems capable of de-coupled, off-line modeling yet not prove susceptible to the kind of functional de-composition involved in a representational analysis. This might be the case if, for example, the inner goings-on are all inter-linked by vast quantities of reciprocal causal influence, thus making it impossible to assign specific representational roles to identifiable sub-states or processes. Under such conditions, the most we might say is that the whole tangled web of inner resources constitutes a kind of miniature dynamical re-construction of the relevant aspects of extra-neural reality. But if we cannot go on to be more specific—if we cannot also assign finer-grained representational roles to component states or processes—then the explanatory value of the analysis is significantly (perhaps fatally) reduced.

This possibility (discussed at length in Clark, 1997a) should not depress us unduly. It is, after all, one of the recurring demands of the sceptic about internal representation that the notion be unpacked in such a way as to make clear its empirical commitments. It is thus a virtue of our account that it exposes the strong representationalist vision to empirical refutation. That said, it is comforting to reflect that on-going neuroscientific research looks to reveal something less than the kind of burgeoning complexity described above: dense reciprocal influences are indeed widespread in the brain, but it still seems possible to identify specific aspects of neural circuitry with specific representational roles—witness e.g., the work on cells in the cat red nucleus mentioned previously. It is also worth remarking that, even where burgeoning complexity threatens, it is sometimes possible to find descriptions that allow us to associate representational roles with higher level systemic features including dynamical constructs such as limit cycles, potential wells, and attractors formed in the spaces defined by complex coupled systems. In these cases, the surface appearance of analytically-intractable complexity hides lower-dimensional patterns that may play specific and identifiable representational roles (Busemeyer & Townsend, 1995; Clark, 1997b).

Notice, finally, that the bulk of contemporary anti-representationalist rhetoric and argument fails to engage with our position since it focuses heavily on the complexities of on-going closed loop agent-environment interactions. Such interactions do, we admit, tend to involve complex, continuous, reciprocal causal influences (the agent is continuously affecting some aspect of the world while that aspect is, simultaneously, continuously influencing the agent. . .) of an analytically problematic kind. But our story is explicitly focused on cases in which cognitive processing proceeds outside of any such closed-loop routines.

4. CONCLUSIONS: WHAT MAKES A SCIENCE COGNITIVE?

Cognitive science, it may safely be presumed, is the science (or sciences) whose object is the explanation and understanding of cognition. But what makes a phenomenon cognitive in the first place? We suggest that the truly cognitive phenomena are those that involve off-line reasoning, vicarious environmental exploration, and the like. It is worth underlining the fact that this stance places us somewhat at odds with an increasingly influential view that either rejects the idea of a cognitive/non-cognitive divide altogether or (more commonly) expands the realm of the cognitive to include all kinds of adaptively valuable organism/environment coupling (Kelso, 1995, pp. 33-34; Thelen & Smith, 1994, pp. 311-339). As two prominent theorists put it:

> "Cognitive processes span the brain, the body, and the environment: to understand cognition is to understand the interplay of all three. Inner reasoning processes are no more essentially cognitive that the skillful execution of coordinated movement or the nature of the environment in which cognition takes place."
> (van Gelder & Port, 1995, p. viii-ix)

We think this is just too liberal to be explanatorily useful. Is not the notion of a truly cognitive agent, at root, the notion of something like a reflective agent? What is needed, we believe, is just a principled way to make this idea (of a reflective agent) precise and to purge it of its original (but probably superficial) associations with episodes of conscious reflection. Our notion of strong representation lets us do just this. Cognizers, on our account, must display the capacity for environmentally decoupled thought and the contemplation of options. The cognizer is thus the being who can think or reason about its world without directly engaging those aspects of the world that its thoughts concern. This is the essence of the emulator example developed in Section 1 and affords (we suggest) an intuitive yet reasonably precise means of demarcating cognitive agency from other (important) kinds of adaptive response. Notice however that this focus on decoupled reason can still accommodate cases in which extra-neural apparatus and structures are playing genuinely cognitive roles e.g., the use of pen and paper to create a kind of extended emulator circuit (Clark & Chalmers, 1997; Hutchins, 1995).

In thus suggesting a substantial demarcation within the class of adaptively potent mechanisms and ploys, we do not mean to in any way disparage the large body of recent work focused on simpler kinds of agent/environment interaction. In fact, we hold that fluent, coupled real-world action-taking is a necessary *component* of cognition. The states of the emulator circuit count as representations in part because of the way they can feed into and affect the system's more basic engagements with its world. Cognition, we want to say, requires both fluent real-world coupling and the capacity to improve such engagements by the use of de-coupled, off-line reasoning.

Two final points should establish the genuine distance between our view and more traditional and restrictive forms of cognitivism and representationalism. First, we recognize the adaptive niche and the real biomechanics of the organism to be crucial determinants of what kinds of states of affairs it may need to (strongly) represent and of how it should represent them. Analog models of egocentrically-specified aspects of the agent's own body and local surroundings are thus perfectly compatible with our brand of representationalism. Second, we do not advocate the focus on top level cases (such as mental arithmetic, chess playing and so on) that characterized early cognitivist research. The motor emulation case was chosen precisely because it displays one way in which the investigation of quite biologically basic, robotics-friendly problems can nonetheless phase directly into the investigation of agents that genuinely cognize their worlds. Such agents are able to substitute inner dynamics for on-going environmental stimulation, and command adaptively valuable inner spaces that they use to sculpt and modulate their more direct engagements with the world. It is these 'Cartesian Agents,' we believe, that must form the proper subject matter of any truly *cognitive* robotics.

NOTES

[1] This section draws on material developed more fully in Grush (1995, 1997).

[2] We are explicitly excluding reflexes, cyclic and automatic movements, such as chewing and walking, from the discussion. Such motor skills may be subserved by mechanisms quite different than ones we discuss here, such as central pattern generators, reflex arcs, etc.

[3] Quantitatively, there looks to be a minimum of at least 200ms, and perhaps as great as 500 ms, delay between onset (at the sensory periphery) of the signal and its actual use in regulating activity of the arm. This figure is established by, for example, using artificial vibrators strapped to the tendons to disrupt the proprioceptive signal from the muscle spindles, and timing the gap between such disruptive input and alterations to the arm motion itself (see Denier van der Gon, 1988; Redon et al., 1991). Yet fine-grained alterations which correct variances in the initial part of a movement appear in normal reaching within the first 70 milliseconds of the action (van der Meulen et al., 1990); (Grush, 1995).

[4] The emulators we describe here are quite simple. In actual engineering applications, these emulators would be a part of a more sophisticated control structure, such as a model-reference adaptive controller or a Kalman filter (see, e.g., Wolpert et al., 1995). This does not effect the points we make, however.

[5] The present discussion of emulator circuitry is necessarily quite short, but additional material filling out the major claims can be found in the following publications: For evidence that proprioceptive feedback is to slow to appropriately aid fast voluntary movements (Denier van der Gon, 1988); (van der Meulen et al., 1990). For models which use emulator circuitry (also known as forward models) in order to solve this problem, see Wolpert et al., 1995; Kawato, 1990. For further discussion of these issues, together with discussion of their relation to imagery, both motor and visual, see Grush 1995, 1997, in preparation.

[6] Interestingly, the authors who originally discovered and published this data (Ghez, 1990); (Ghez & Vicario,

1978) interpret the function of these neurons quite differently than we do. They see them as different classes of motor neurons, which just happen (for reasons unexplained) to closely mimic the physical parameters. This interpretation and ours are not incompatible, however.

[7] Notice then, that our concern is not with whether the systems display 'intrinsic intentionality' or have real thoughts or emotions. We intend merely to distinguish two ways of supporting successful behavior; one of which involves de-coupleable inner stand-ins while the other does not.

[8] As a matter of record, one of us, (Grush), is not sure he agrees with this and is drawn to the stronger claim that, in fact, strong representation is a necessary condition for cognitive status.

REFERENCES

Agre, P. (1996). Computational research on interaction and agency. In P. Agre & S. Rosenschein (Eds.), *Computational Theories of Interaction and Agency*. Cambridge, MA: MIT Press.

Agre, P., & Rosenschein, S. (Eds.). (1996). *Computational Theories of Interaction & Agency*. Cambridge, MA: MIT Press.

Beer, R. (1995). A Dynamical Systems Perspective on Agent-Environment Interaction. *Artificial Intelligence, 72*, 173-215.

Brooks, R. (1994). Coherent Behavior from Many Adaptive Processes. In D. Cliff, P. Hubands, J. A. Meyer, & S. Wilson (Eds.), *From Animals to Animats 3* (pp. 22-29). Cambridge, MA: MIT Press.

Brooks, R., & Stein, L. (1993). *Building Brains for Bodies* (Memo 1439): Artificial Intelligence Laboratory, Cambridge, MA.

Busemeyer, J., & Townsend, J. (1995). Decision Field Theory. In R. Port & T. Van Gelder (Eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Campbell, D. (1974). Evolutionary Epistemology. In P. Schilpp (Ed.), *The Philosophy of Karl Popper* (pp. 413-463). LaSalle, IL: Open Court.

Churchland, P., & Sejnowski, T. (1992). *The Computational Brain*. Cambridge, MA: MIT Press.

Clark, A. (1997a). *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Clark, A. (1997b) The Dynamical Challenge. *Cognitive Science*, *21*(4), 461-481.

Clark, A. and Chalmers, D. (1997). The Extended Mind. *Analysis, 58:* 7-19.

Clark, A., & Karmiloff-Smith, A. (1993). The Cognizer's Innards: A Psychological & Philosophical Perspective on the Development of Thought. *Mind & Language, 4:* 487-519.

Dean, P., Mayhew, J., & Langdon, P. (1994). Learning and Maintaining Saccadic Accuracy: A Model of Brainstem-Cerebellar Interactions. *Journal of Cognitive Neuroscience, 6*(2), 117-138.

Dennett, D. (1995). *Darwin's Dangerous Idea*. New York: Simon & Schuster.

Denier van der Gon, J.J. (1988). *Motor Control: Aspects of its Organization, Control Signals and Properties.* Paper presented at the Seventh Congress of the International Electrophysiological Society.

Elman, J. (1993). Learning and Development in Neural Networks: The Importance of Starting Small. *Cognition, 48*, 71-99.

Fodor, J. (1975). *The Language of Thought*. New York: Crowell.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge: MIT Press.

Ghez, C. (1990). Voluntary Movement. In Kandel, E., Schwartz, J., and Jessell, T. eds. *Principles of Neural Science*, Third Edition. Elsevier.

Ghez, C. and Vicaro, D. (1978). Discharge of red nucleus neurons during voluntary muscle contraction: activity patterns and correlations with isometric force. *Journal of Physiology, Paris 74*, 283-285.

Grush, R. (1995). *Emulation & Cognition*. Unpublished Ph.D. Dissertation, University of California, San Diego.

Grush, R. (1997). The Architecture of Representation. *Philosophical Psychology, 10*(1), 5-23.

Grush, R. (in preparation). *The Machinery of Mindedness.*

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* Cambridge: MIT Press.

Haugeland, J. (1991). Semantic Engines: An Introduction to Mind Design. In J. Haugeland (Ed.), *Mind Design: Philosophy, Psychology, Artificial Intelligence* (pp. 1-34). Cambridge, MA: MIT Press.

Hutchins, E. (1995) *Cognition in the wild.* Cambridge, MA: MIT Press.

Israel, D. (1988). Commentary: Bogdan on Information. *Mind & Language, 3*(2), 123-140.

Ito, M. (1984). *The Cerebellum & Neural Control.* New York: Raven Press.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science.* Cambridge, MA: MIT Press.

Kawato, M. (1990). Computational schemes and neural network models for formation and control of multijoint arm trajectory. In W. T. Miller III, R. Sutton, & P. Werbos (Eds.), *Neural Networks for Control* . Cambridge, MA: MIT Press.

Kawato, M., Furukawa, K., & Suzuki, R. (1987). A hierarchical neural network model for the control and learning of voluntary movement. *Biological Cybernetics, 57,* 169-185.

Kelso, S. (1995). *Dynamic Patterns*. Cambridge, MA: MIT Press.

McClelland, J., Rumelhart, D., & Group, P. R. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. (Vol. I & II).* Cambridge, MA: MIT Press.

Port, R., & Gelder, T. V. (Eds.). (1995). *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.

Pylyshyn, Z. (Ed.). (1987). *The Robots Dilemma: The Frame Problem in Artificial Intelligence.* Norwood: Ablex.

Redon, Christine, Hay, Laurette, and Velay, Jean-Luc (1991). Proprioceptive control of goal-directed movements in man, studied by means of vibratory muscle tendon stimulation. *Journal of Motor Behavior 23*(2), 101-108.

Smith, B. C. (1996). *On the Origin of Objects*. Cambridge, MA: MIT Press.

Smolensky, P. (1988). On the Proper Treatment of Connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Stufflebeam, R. (1996). *Whither Internal Representations?* Unpublished Ph.D. Dissertation, Washington University, St. Louis.

Thelen, E., & Smith, L. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action.* Cambridge, MA: MIT Press.

Touretsky, D. & Pomerleau D. (1994). Reconstructing Physical Symbol Systems. *Cognitive Science, 18*, 345-353.

Ungar, L. (1990). A bioreactor benchmark for adaptive network-based process control'. In W. Miller, R. Sutton, & P. Werbos (Eds.), *Neural Networks for Control.* Cambridge, MA: MIT Press.

van der Meulen, J.H.P., Gooskens, R.H.J.M., Denier van der Gon, J.J., Gielen, C.C.A.M., and Wilhelm, K. (1990). Mechanisms underlying accuracy in fast goal-directed arm movements in man. *Journal of Motor Behavior, 22*(1), 67-84.

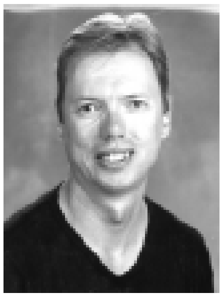
Van Gelder, T. (1995). What Might Cognition Be, If Not Computation? *Journal of Philosophy, XCII*(7), 345-381.

Van Gelder, Timothy, and Port, Robert (1995) It's About Time: An Overview of the Dynamical Approach to Cognition. In Port and van Gelder, eds. (1995). *Mind as Motion: Explorations in the Dynamics of Cognition.* Cambridge, MA: MIT Press.

Vera, A., & Simon, H. (1993). Situated Action: A Symbolic Interpretation. *Cognitive Science, 17*, 4-48.

Wheeler, M. (1994). From Activation to Activity. *Artificial Intelligence and the Simulation of Behavior (AISB) Quarterly, 87*, 36-42.

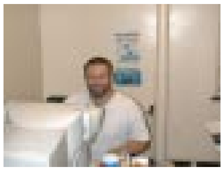
ABOUT THE AUTHORS



Andy Clark

Andy Clark is Professor of Philosophy and Director of the Philosophy/Neuroscience/Psychology program at Washington University in St. Louis, Missouri. Educated at the University of Stirling, Scotland, Clark pursued undergraduate studies in Philosophy and took a Ph.D. addressing issues in Philosophy and Evolutionary Biology. He taught briefly at the University of Glasgow, Scotland before accepting a "New Blood" appointment at the University of Sussex, England in 1985 where he later became Reader and then Professor in Philosophy and Cognitive Sciences. He is the author of three books Microcognition (MIT Press/Bradford Books 1989), Associative Engines (MIT Press/Bradford Books, 1993), and Being There: Putting Brain, Body and World Together Again (MIT Press 1997).

Rick Grush



Rick Grush is currently Assistant Professor in the Philosophy Department at the University of Pittsburgh. His PhD is in Cognitive Science and Philosophy, from UC San Diego, and as awarded in 1995. He was McDonnell Postdoctoral Fellow at the Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis from 1995 - 1997, and held a research position at the Center for Semiotic Research at the Unversity of Aarhus, Denmark, from 1997 - 1998. His research is organized around the attempt to arrive at an understanding of how neural systems can be thinking, representing, epistemically reflective, language-using systems. More specifically, he has strong current research interests in the work of the philosopher Gareth Evans, the neurobiology of spatial representation, and cognitive linguistics, especially Ronald Langacker's Cognitive Grammar framework. He has articles pubished in a number of journals, including *Philosophical Psychology*, and *The Journal of Consciousness Studies* (with Patricia Churchland), as well as a number of invited contributions to various books. His current projects include guest-editing a special issue of *The Electronic Journal of Analytic Philosophy* devoted to the work of Gareth Evans, and a book manuscript, currently titled *The Machinery of Mindedness*.