MIND AND LANGUAGE 16:2: 2001 p.121-145

Reasons, Robots and the Extended Mind

(Rationality for the New Millenium)*

Andy Clark
School of Cognitive and Computing Sciences
University of Sussex
Falmer
Brighton
BN1 9QH

andycl@cogs.susx.ac.uk

* Special thanks to Sam Guttenplan for advice, criticism and editorial guidance.

Abstract

A suitable project for the new Millenium is to radically reconfigure our image of human rationality. Such a project is already underway, within the Cognitive Sciences, under the umbrellas of work in Situated Cognition, Distributed and De-centralized Cogition, Real-world Robotics and Artificial Life[1]. Such approaches, however, are often criticized for giving certain aspects of rationality too wide a berth. They focus their attention on on such superficially poor cousins as "adaptive behaviour", "ecologically sound perception-action routines", "fast and frugal heuristics" and "fast, fluent real-time real-world action control"[2]. Is this robbery or revelation? Has 'embodied, embedded' cognitive science simply lost sight of the very phenomena it was meant to explain? Or are we finally seeing rationality aright, as fully continous with various forms of simpler, ecologically situated, adaptive response?

I distinguish two ways of developing the 'embodied, embedded' approach. The first, which does indeed threaten to lose sight of the key targets, is fully committed to a doctrine of *biological cognitive incrementalism* according to which full-scale human rationality is reached, rather directly, by some series of tweaks to basic biological modes of adaptive response. The second depicts human capacities for advanced reason as at best the indirect products of such a process. Such capacities, it is argued, depend heavily upon the effects of a special kind of hybridization in which human brains enter into an increasingly potent cascade of genuinely symbiotic relationships with knowledge-rich artifacts and technologies. This latter approach, I finally suggest, does better justice to our peculiar profile, which combines deep biological continuity with an equally deep cognitive discontinuity.

---

1. For an overview of all these developments, see Clark (1997a)

2 See, for example Brooks (1991), Beer (1995), Gigerenzer and Todd (1999) and Clark (1997a).

1. Introduction: Where the Rubber Meets the Road

The changing of the Millenium is a time to sit back, to take stock, to reflect on the big picture and to ask ourselves "Where to next?". For philosophes of mind, and especially for those of us working within a broadly naturalistic framework, that means reflecting on (amongst other things) the shape and scope of scientific explanations of human reason. For human rationality, without a doubt, is where the rubber meets the road for traditional Cognitive Scientific approaches to understanding mind and reason. Putting aside any purely (and problematically) normative questions (such as "how ought we to reason?"), philosophers and cognitive scientists have found common cause in the attempt to understand how such rationality as we in fact display is, as Jerry Fodor likes to put it, "mechanically possible" (Fodor (1998) p.204 and elsewhere).

Yet withing the Cognitive Sciences, there stirs a strangeness. Many of the most exciting recent research programs give traditional visions of rationality a fairly wide berth, focussing their efforts and attention on such superficially poor cousins as “adaptive behaviour”, “ecologically sound perception-action routines”, “fast and frugal heuristics” and “fast, fluent real-time, real-world action control”(see note 2). Is this robbery or revelation? Has 'embodied, embedded' cognitive science simply lost sight of the very phenomena it was meant to explain? Or are we finally seeing rationality aright, as fully continous with various forms of simpler, ecologically situated, adaptive response?

I distinguish two ways of developing the 'embodied, embedded' approach. The first, which does indeed threaten to lose sight of the key targets, is fully committed to a doctrine of *biological cognitive incrementalism* according to which full-scale human rationality is reached, rather directly, by some series of tweaks to basic biological modes of adaptive response. The second depicts human capacities for advanced reason as at best the indirect products of such a process. Such capacities, it is argued, depend heavily upon the effects of a special kind of hybridization in which human brains enter into an increasingly potent cascade of genuinely symbiotic relationships with knowledge-rich artifacts and

technologies. This latter approach, I finally suggest, does better justice to our peculiar profile, which combines deep biological continuity with an equally deep cognitive discontinuity. Recognizing and understanding this dual profile, I shall argue, is an important step towards reconciling the insights of the traditional and the "embodied, embedded" camps. What follows is at best a preliminary sketch of this proposed reconciliation. The hope, in line with the millenial brief, is simply to display a growing tension, and to scout a few ways to move ahead.

I begin, then, with a brief rehearsal of some twice-told stories about the mechanical roots of human reason.

2. Mechanical Models of *What*?[3]

A Jerry Fodor printbite sets the scene. "Beyond any doubt" Fodor lately assures us:

> the most important thing that has happened in cognitive science was Turing's invention of the notion of mechanical rationality' (Fodor (1998) p.204).

Cognitive science, following through on Turing's vision, is thus to provide us with a mechanical model of something variously called "reason" or "rationality" or (sometimes) "thinking".And early work in cognitive science was indeed dominated, as Fodor suggests, by a Turing-machine inspired vision of reading, writing, and transposing chunky symbols. Typical operations included copying, combining, creating and destroying text-like symbols according to instructions. Such accounts excelled in explaining how simple, sententially-couched inferences might be mechanically reproduced. It explained them by pairing each participating thought (as expressed in words) with an inner symbolic echo sharing relevant aspects of the structure of the putative thought. Simple syntax-

---

[3]With apologies to Ian Hacking (*The Social Construction of What?* , Harvard University Press, Cambridge, MA, 1999).

sensitive computational processes could then regulate local inferences in ways that marched in step with certain semantic relations between the thoughts - truth-preservingness here being the simplest example (see e.g. Newell and Simon (1981), Haugeland (1981, 1997)). It is with this kind of story in mind that Fodor goes on to comment that:

> Some, at least, of what makes minds rational is their ability to perform computations on thoughts; where thoughts, like sentences, are assumed to be syntactically structured, and where "computations" means formal operations in the manner of Turing. (Fodor (1998) p.205).

(The 'some' will later prove important. Fodor, it turns out, is way more circumspect concerning exactly what the Turing Machine conception can explain than one might sometimes imagine).

This image of inner symbols in motion has, of course, a famous competitor. This is the "connectionist" vision of reason-guided thought transitions as grounded in vector-to-vector transformations within a large web of simple, parallel computing elements. I shall not attempt a proper discussion of this alternative here (but see Clark (1989), (1993) for some attempts). A noteworthy point of contrast, however, concerns the "best targets" for each approach. For while classical approaches excelled at modelling rational inferences that could be displayed in *sentential space*, connectionist work excelled at modelling those dimensions of rationality best conceived of as *skill-based*. By "sentential space" I mean an abstract space populated by meaning-carrying items that share the logical form of sentences: sequential strings of content-bearing elements in which different syntactic items reliably stand for different things and in which the overall meaning is a function of the items and their sequential order. By "skill-based dimensions of rationality" I mean the reliable capacity to perform "inferences" in which the inputs are, broadly speaking, perceptual and the outputs are, broadly speaking, motoric. Examples of the former (sentential-type) include the inference from 'it's raining' and 'I hate getting wet' to 'I'd better take an umbrella'. Examples of the latter (skill-based) include the cat's rapid assessment of the load bearing capacity of a branch, leading to a swift and elegant leap. Of course, inference-making in sentential space can involve

perceptual inputs and motoric outputs (such as speech) too. And connectionists tend, indeed, to view the sentential inferences as just special cases of their more general skill-based vision (see P.M.Churchland 1995). Making an expert medical judgement, for the connectionist, has more in common with knowing how to ride a bicycle than with consulting a set of rules in a symbolic data-base (Dreyfus and Dreyfus (1990)). Reasoning and inference are thus reconstructed, *on all levels* as (roughly speaking) processes of pattern-evolution and pattern-completion carried out by cascades of vector-to-vector transformations in parallel populations of simple processing units.

The most recent waves of work in cognitive science continue this process of inner-symbol flight, moving us even further from the traditional shores of linguiform reason. Most of this work (as we'll see more clearly in the next section) has very little - or at any rate, very little *positive* - to say about the traditional targets. Instead, we find accounts of the neural control dynamics for insect-like robot bodies (Beer (1995)), of the interplay between leg mass and spring-like action in the development of infant stepping (Thelen and Smith (1994)), of the complex dynamics of co-ordinated finger wiggling (Kelso (1995)), of the rich contributions of bodily action and task environments to visual processing (Ballard (1991)), and in general - to conclude a potentially much more extensive list - of the many and various ways body, brain and world interact so as to support fast, fluent responses in ecologically normal settings (see Clark (1997a) for a review). So whatever happened to reason, thought and rationality?

Two things happened, I think. The first is that folk began to doubt the *centrality* of the traditional exemplars of reason, thought and rationality. Take, for example, Jerry Fodor's favourite inference: 'It's raining, I don't want to get wet, therefore I shall take my umbrella' (e.g. Fodor (1987) p.17). Such inferences, while perhaps more important in New York than in, say, San Diego, seem somewhat less than paradigmatic of the kind of thing upon which survival and reproduction most centrally turns. A better bet might be the "inference" from 'object looming ever larger in my visual field' to 'something is moving very fast towards me' to 'let's duck'. But *this* kind of chain of effect, it has long been clear, is generally underwritten by mechanisms whose form and functioning is quite unlike those originally described by Turing, and celebrated by Fodor and others. A typical story might, for example, associate the ducking response with the visually mediated detection of a rather simple perceptual invariant (computed,

without any need for shape information, from the raw outputs of directionally sensitive cells in two adjacent retinas (see Marr (1982)). Similar kinds of reflex-like stories account for e.g. the well-timed wing closing of diving birds, such as gannets (see Lee and Reddish (1981), Gibson (1979)).

Such stories stand in sharp contrast to ones in which the relevant premisses are literally encoded in an inner language, and the inferences conducted by well regulated processes of symbol manipulation and transformation, terminating in an expression (in that same code) of a possible and appropriate action. It is unsurprising, then, that the kinds of capacity and behaviour targeted by the most successful "alternative" stories are (intuitively speaking) less "cognitive" and more "automatic" than those targeted by many traditional accounts: looming-detection, object ducking, walking and wall-following on the one hand; conversation, story-understanding and means-end reasoning on the other[4]. The possibility thus arises that the two research programs are simply talking past each other - appearing to disagree about the mechanistic underpinnings of mind while in fact targeting quite different dimensions of the same system. But such a comforting diagnosis is too quick. For the real point of the alternative accounts is indeed to call into question the worth and centrality of many of the traditional exemplars themselves: to call into question the idea that the heart of human reason lies in the mechanisms that support inferences such as 'it's raining, so I'd better take my umbrella'.

Combined with this doubt about the centrality of the traditional exemplars we find a second motivating idea: the idea that to avoid distorting our science we should *creep up* on the high-level, distinctively human kinds of thought and reason by a process of incremental investigation which begins by targeting much more basic aspects of adaptive response (such as obstacle avoidance, predator recognition and so on). This, then, is the *second* thing that has happened to naturalistic investigations of reason, thought and rationality. Overall, the new sciences of the mind have become deeply coloured by an underlying belief in what I'll call "biological cognitive incrementalism" - the idea that there is a fairly smooth sequence of tweaks and small design innovations

[4] These are just examples of course. For a better sense of the range of phenomena targeted by this recent work, see Clark (1997a), and for a sense of the more traditional targets, see the essays by Minsky, Dreyfus and Newell and Simon, in Haugeland (1981).

linking full-scale human thought and reason to the mechanisms and strategies deployed in more basic, and less intuitively reason-sensitive, cases.

These doubts about centrality, and the supporting belief in biological cognitive incrementalism, are not, of course, entirely new. Both Daniel Dennett and Paul Churchland have, in different ways, long doubted the centrality of the kinds of thinking highlighted by traditional accounts (see e.g. Dennett (1982), and the comments on page 363 of his (1998), or Churchland's (1989) (1995) marginalization of sentence-like encodings). And David Marr, as long ago as 1981, was clearly committed to something like the story about cognitive incrementalism (see Marr (1981) p.140). What is new, however, is the sheer range of tools and models being developed to target these more basic forms of adaptive success, the widespread popularity of the project, and the increasingly radical theoretical agenda underlying much of the work. It is an agenda which is nicely summed up by Godfrey-Smith in his (1996, p320) description of a thesis of 'strong continuity' according to which;

> Life and mind have a common abstract pattern or set of basic organizational properties. The functional properties characteristic of mind are an enriched version of the functional properties that are fundamental to life in general. Mind is literally life-*like*[5].

Connectionism already took one step in this direction, assimilating the processes underlying *reason* to those essential to most forms of skilled adaptive response. Recent work in robotics and Artificial Life (see below) takes this process several steps further, as we shall now see.

2. Understanding a Robot Cricket

[5] As far as I can tell, Godfrey-Smith himself remains agnostic on the truth of the strong continuity thesis. He merely presents it as one of several possible positions and relates it to certain trends in the history of ideas (see Godfrey-Smith (1996)).

Recent bodies of work in 'embodied, embedded'[6] cognitive science constitute a quite radical departure from the Turing-Fodor vision, and even - in some ways - from the connectionist skill-based vision. In much of this recent work, traditional conceptions of thought, reason and action are not so much re-worked as by-passed entirely. I have discussed many of these developments elsewhere (especially Clark (1997a), (1999)) and must here be content with a rather summary sketch.

Three notable features of the new work are:

1. The attempt to model and understand *tightly coupled* organism-environment interaction;

2. An increasing recognition of the role of *multiple, locally effective tricks, heuristics and short-cuts* (often action-involving) in supporting adaptive success,

and

3. Attention to various forms of *socially, culturally and technologically distributed problem-solving* allowing the "off-loading" of substantial problem-solving activity into the local environment.

The overall effect of this three-pronged assault has been to re-invent rationality as an active, distributed, environment-exploiting achievement. Over-attention to the first prong has, however, caused some theorists to almost lose sight of the traditional targets. Thus, at the most radical end of the spectrum of new approaches we find deep scepticism about the important piece of common-ground shared by classicists and connectionists alike: the vision of rational action as involving the deployment and transformation of internal representations.

---

[6] The phrase 'embodied, embedded cognitive science' is is due to Haugeland, and can be found in the essay reprinted as Haugeland (1998)

To get the flavour of this scepticism, consider the humble house-fly. The fly, as Marr (1982, p.32-33)[7] points out, gets by without needing to know that the act of flying requires the flapping of wings. For circuitry running from the fly's feet ensures that, whenever the feet are not in contact with a surface, the wings flap. All the fly has to do is jump off the wall, and the flapping will follow. But now imagine such dumb circuitry multiplied. Suppose the jump routine is itself automatically activated whenever a certain ('looming shadow indicating') perceptual invariant is detected in the raw visual input. Now let's go the whole hog (or is it the whole Iguana? See Dennett (1978)) and imagine a complete creature composed entirely of such routines, all turning each other on and off at what are (generally speaking) ecologically appropriate moments. What you have imagined is - coarsely but not, I think, too inaccurately - the kind of "subsumption architecture" introduced by the roboticist Rodney Brooks and discussed in provocatively named papers such as "Intelligence without Representation" (Brooks (1991)).

Brooks himself, it is now apparent, was concerned mainly to avoid (what he saw as) the excesses of traditional symbolic approaches. He allows that we may choose to treat the states of certain specific circuits, even in a subsumption architecture, as 'representing' such-and-such. What we should not do (or not do too soon) is to posit a richly expressive inner language, complicated mechanisms of search and inference, and an inner economy that trades heavily in messages couched in the inner language. The extent of Brooks' opposition to the connectionist alternative is less clear and probably depends on the precise details of any broadly connectionist model. Brooks (op cit) does, however, explicitly note that the subsumption architecture idea is not just connectionism re-heated, but aims instead for a distinctive, and especially representation-sparse, kind of design.

Or consider Barbara Webb's wonderful work on phonotaxis in a robot cricket. Phonotaxis is the process whereby a female cricket identifies a male of the same species by his song, turns in his direction, and reliably locomotes to the source. On the face of it, this looks to be a complex, three part problem. First,

---

[7] This nice example is reported in McClamrock (1995, p.85).

hear various songs and identify the specific song of the male of your own species. Second, localize the source of the song. Third, locomote that way (making needed corrections en route). This way of posing the problem also makes it look as if what the cricket will require to solve the problem are some quite general cognitive capacities (to recognize a sound, to discern it's source, to plot a route). Nature, however is much thriftier, and has gifted the cricket with a single, efficient, but totally special-purpose problem-solving procedure.

The cricket's ears are on its forelegs and are joined by an inner tracheal tube that also opens to the world at two points (called spiracles) on the body. External sounds thus arrive at each ear by two routes: directly (sound source to ear) and indirectly (via the other ear, spiracles and tracheal tube). These two routes take significantly different amounts of time on the side nearest the sound source, where the direct route is much quicker than the indirect one. The extra time taken to travel through the tube alters the phase of the 'inner route' sound relative to the 'outer route' sound on the side (ear) nearest the source. At the eardrum these out of phase sound waves are summed, and the resulting vibration is heard, by the cricket, as a markedly louder sound on the side nearest the source. A dedicated interneuron (one per ear) detects this effect and the cricket turns, automatically, in the direction of the source. This whole system works only because first, the cricket's tracheal tube is especially designed to transmit songs of the species-specific frequency, and because the male song repeats, offering frequent bursts each of which powers one of these episodes of orienting-and-locomoting (hence allowing course corrections, and longer distance mate-finding)- see Webb (1996) for full details.

In the very specific environmental and bio-mechanical matrix that Webb describes, the cricket thus solves the whole (apparently three-part) problem using a single special-purpose system. There is no need, for example, to actively discriminate the song of your own species, because the specifics of your auditory system are structurally incapable of generating the directional response to other sounds. Nor, of course, do you bother to build a model of your local surroundings so as to plan a route. Instead, you (the cricket) exploit neat tricks, heuristics, and features of your body and world. Moreover you (and your real-world sisters) seem to succeed without relying on anything really worth calling internal representations. Webb allows that various inner states do, to be sure,

correspond to salient outer parameters, and various inner variables to specific motor outputs. But:

> 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
>
> Webb (1994), p. 53.

Understanding the behaviour of the robot cricket requires attention to features that, from the standpoint of classical cognitive science, look more like details of implementation (the fixed length trachea, the details of ear and spiracle placement) and environmental context (the syllable repetition rate of the male) than substantive features of an intelligent control system.

Such work exemplifies the recent focus on tightly-coupled organism-environment interactions. As such couplings increase in number and complexity, the already-thin leverage provided by traditional theoretical notions of representation, symbol, and algorithm continues to diminish. Scott Kelso's work on rythmic finger motion (Kelso 1995) and Randy Beer's (1995) work on cockroach locomotion, for example, each lay great stress on the precise timing of specific bio-mechanical events and on the kinds of continuous, circular causal chains in which components X and Y (etc) each continually influence, and are influenced by, the activity of the other. The theoreretical tools of choice, for understanding the behavior of such complex, coupled, temporally-rich systems is, many researchers[8] now argue, dynamical systems theory. Dynamical systems theory is a well-established framework in the physical (not the informational, computational or representational) sciences, geared to modeling and describing phenomena (including tightly coupled systems) that change over time (for discussion, see Clark (1997a, 1997b)).

Attention to new types of target (tightly coupled organism-environment interactions) thus rapidly breeds scepticism concerning the explanatory power of

---

[8] Eg Van Gelder (1995), Thelen and Smith (1994), Kelso (1995), Beer (1995)

the old notions (representation, rule-based search etc): a scepticism which is then extended to a much wider variety of cases. The suspicion (see e.g. Thelen and Smith (1994) Van Gelder (1995), Keijzer (1998)) becomes that symbols, internal representations and the like play little role even in advanced human problem-solving. The trouble with this, as I argue in detail elsewhere (Clark (1997) chapter 8, Clark and Toribio (1994), Clark (1997b)), is that it is not *at all* obvious -indeed, it seems highly implausible - that truly representation-sparse internal architectures can indeed support the kinds of reason-governed, rational behaviour characteristic of our species[9]. Internal representations look prima facie essential for such mundane (to us) activities such as dreaming of Paris, mulling over U.S.gun control policy, planning next years vacation, counting the windows of your New York apartment while on holiday on Rome, and so on. These kinds of behaviour are all (to use a term from Clark and Toribio (1994)) "representation-hungry". All these cases, on the face of it, require the brain to use internal *stand-ins* for potentially absent, abstract, or non-existent states of affairs. A "stand-in", in this strong sense is an item designed not just to carry information about some state of affairs (in the way that e.g., some inner circuit might carry information about the breaking of foot-surface contact in the fly) but to allow the system to key its behaviour to features of specific states of affairs *even in the absence of direct physical connection*. (For some excellent discussion of the topics of connection and disconnection, see B. C. Smith (1996)). By contrast, nearly all[10] the cases typically invoked to show representation-free adaptive response are cases in which the relevant behaviour is continuously driven by, and modified by, ambient input from the states of affairs to which the behaviour is keyed.

I present this, let it be said, not as any kind of knock-down argument. It is still possible that representation-sparse models will account for the full gamut of human activity. But the burden of proof remains, surely, with the sceptics. Displaying representation-sparse success in domains which were not representation-hungry in the first place cannot count as much of a strike against representationalism.

---

[9] The debate is complicated, of course, by the absence of an agreed account of just when some inner state should properly *count* as a representation. See Clark (1997b), Clark and Grush (1999) for a proposal

[10] But see Stein (1994), Beer (2000) and discussion in Clark (1999)

There are two other elements, however, in the embodied, embedded world-view which may ultimately (or so I believe) prove to be of more general importance. One is the emphasis on simple heuristics. There is a growing body of recent work (see especially Gigerenzer, Todd and the ABC Research Group (1999)) that displays the remarkable efficacy of special-purpose routines and heuristics even in 'advanced' human problem-solving. Such procedures, the authors show, can often be shown to perform as well (or sometimes better) as more traditionally 'rational' and knowledge-intensive procedures. Indeed, Gigerenzer et al dispute the idea that the use of (what they call) 'fast and frugal heuristics' amounts to some kind of failure of rationality and instead depict such strategies as the heart and soul of 'ecological rationality'. Thus although there is a long tradition in A.I and psychology (such as the 'heursitics-and-biases program- see Tversky and Kahneman (1974)) of studying 'quick-and-dirty' short-cuts, the new research does so in a rather different spirit:

> Whereas the heuristics-and-biases program portrays heuristics as a frequent hindrance to sound reasoning, rendering *Homo sapiens* not so sapient, we see fast and frugal heuristics as enabling us to make reasonable decisions and behave adaptively in our environment- *Homo sapiens* would be lost without them
>
> Gigerenzer et al (1999) p.29

What makes these fast and frugal heuristics *work*, the authors note, is primarily the way they enable the mind to make maximal use of 'the structure of information in the environment' (op cit p.28). Gigerenzer et al note, towards the end of the book, that recent work in robotics (they cite Brooks especially) can be seen as operating in much the same spirit. This can be seen quite clearly, I think, in the specific example of cricket phonotaxis discussed above. But whereas Brooks and others are suspicious of the very idea of mental inference and internal model-building, Gigerenzer et al simply seek to display such activities as computationally tractable and as rather tightly geared to specific structures of information and opportunity in the local problem-solving environment. The second prong of the embodied, embedded approach thus shades into the third: the exploitation of various aspects of the environment as integral parts of the

problem-solving engine. It is this third prong, however, on which I shall now concentrate. Before doing so, however, let's pause to take stock.

We began with the Turing-Fodor vision of reading, writing and transposing inner, language-like symbols. We touched briefly on the delights of connectionist vector coding as a skill-based alternative to the classical vision. Connectionism, we might say, taught us how to trade in *internal representations* without the use of fully-fledged, text-like *inner symbols.* And we ended by reviewing recent work in robotics, artificial life and dynamical systems theory: work which targets the complexities of real-world, real-time organism-environment interactions. This work is, in many ways, the natural culmination of the process (a kind of 'inner symbol flight' -see Clark (in Press)) which the connectionists began.

But what *exactly* does this work have to teach us about full-scale human rationality? Does it simply miss this target altogether? Is the embodied, environmentally embedded approach to cognitive science robbery or revelation? It is robbery, I suggested, if the lesson is meant to be that we can now dispense with the ideas of internal representation and model-building altogether. But it may be revelation if instead it draws attention to the many, complex and underappreciated ways in which minds make the most of the environments they inhabit and (in our case) repeatedly design and build.

3. World, Technology and Reason

Much of advanced cognition, I believe, depends on the operation of the same basic kinds of capacity used for on-line, adaptive response, but tuned and applied to the very special domain of *external and/or artificial cognitive aids* – the domain, as I shall say, of *cognitive technology.* This idea takes its cue from Dennett (1995) (1996), Hutchins (1995), Kirsh and Maglio (1994) and many others. The central idea is that understanding what is *distinctive* about human thought and reason may turn out to depend on a much broader focus than that to which cognitive science has become most accustomed: a focus that includes not just body, brain and the natural world, but the technological props, aids and scaffoldings (pens, papers, PC's, institutions...) in which our biological brains learn, mature and operate.

Work on simpler systems such as the robot cricket already draws our attention to the complex ways in which neural and non-neural (in this case bodily and environmental) factors and forces may combine to render complex problems computationally tractable. The most promising route for understanding and modelling human-level intelligence, I suggest, is to take quite seriously the idea of complex internal representations (though their form may be connectionist rather than classical) but to combine this with a special emphasis on the peculier ways in which human thought and reason is sculpted, enhanced,and ultimately transformed, by our interactions with the rather special environments we create.

That we humans benefit deeply from the empowering presence of non-biological props and scaffolds is, I suppose, indisputable. But we are often blind to the sheer variety, heterogeneity and depth of the non-biological contributions. We tend, mistakenly, to view the main role of these non-biological systems as that of providing more (and more durable, and shareable) *memory.* The real contribution of the props and aids is, however, more like the provision of brand-new computer functionality by the insertion of a PC card than the simple provision of more memory or new storage media.

Thus consider two brief examples: one old (see Clark (1997a) Epilogue) and one new. The old one first. Take the familiar process of writing an academic paper. Confronted, at last, with the shiny finished product the good materialist may find herself congratulating her brain on its good work. But this is misleading. It is misleading not simply because (as usual) most of the ideas were not our own anyway, but because the structure, form and flow of the final product often depends heavily on the complex ways the brain co-operates with, and depends on, various special features of the media and technologies with which it continually interacts. We tend to think of our biological brains as the point source of the whole final content. But if we look a little more closely what we may often find is that the biological brain participated in some potent and iterated loops through the cognitive technological environment. We began, perhaps, by looking over some old notes, then turned to some original sources. As we read, our brain generated a few fragmentary, on-the-spot responses which were duly stored as marks on the page, or in the margins. This cycle repeats,

pausing to loop back to the original plans and sketches, amending them in the same fragmentary, on-the-spot fashion. This whole process of critiquing, re-arranging ,streamlining and linking is deeply informed by quite specific properties of the external media, which allow the sequence of simple reactions to become organized and grow (hopefully) into something like an argument. The brain's role is crucial and special. But it is not the whole story. In fact, the true (fast and frugal!) power and beauty of the brain's role is that it acts as a mediating factor in a variety of complex and iterated processes which continually loop between brain, body and technological environment. And it is this larger system which solves the problem. We thus confront the cognitive equivalent of Dawkins' (1982) vision of the extended phenotype. The intelligent process just *is* the spatially and temporally extended one which zig-zags between brain, body and world.

Or consider, to take a superficially very different kind of case, the role of sketching in certain processes of artistic creation. Van Leeuwen, Verstijnen and Hekkert (1999) offer a careful account of the creation of certain forms of abstract art, depicting such creation as heavily dependent upon "an interactive process of imagining, sketching and evaluating [then re-sketching, re-evaluating, etc.] " (op cit p. 180). The question the authors pursue is: why the need to sketch? Why not simply imagine the final artwork "in the mind's eye" and then execute it directly on the canvas? The answer they develop, in great detail and using multiple real case-studies, is that human thought is constrained, in mental imagery, in some very specific ways in which it is *not* constrained during on-line perception. In particular, our mental images seem to be more interpretively fixed: less able to reveal novel forms and components. Suggestive evidence for such constraints includes the intriguing demonstration (Chambers and Reisberg (1989)) that it is much harder to discover (for the first time) the second interpretation of an ambiguous figure (such as the duck/rabbit) in recall and imagination than when confronted with a real drawing. Good imagers, who proved unable to discover a second interpretation in the mind's eye, were able nonetheless to draw what they had seen from memory and, by then perceptually inspecting their own unaided drawing, to find the second interpretation. Certain forms of abstract art, Van Leeuwen et al go on to argue, likewise, depend heavily on the deliberate creation of "multi-layered meanings" – cases where a visual form, on continued

inspection, supports multiple different structural interpretations. Given the postulated constraints on mental imagery, it is likely that the discovery of such multiply interpretable forms will depend heavily on the kind of trial and error process in which we first sketch and then perceptually (not merely imaginatively) re-encounter visual forms, which we can then tweak and re-sketch so as to create a product that supports an increasingly multi-layered set of structural interpretations. This description of artistic creativity is strikingly similar, it seems to me, to our story about academic creativity. The sketch-pad is not just a convenience for the artist, nor simply a kind of external memory or durable medium for the storage of particular ideas. Instead, the iterated process of externalizing and re-perceiving is integral to the process of artistic cognition itself.

One useful way to understand the cognitive role of many of our self-created cognitive technologies is thus as affording *complementary* operations to those that come most naturally to biological brains. Recall the connectionist image of biological brains as pattern-completing engines. Such devices are adept at linking patterns of current sensory input with associated information: you hear the first bars of the song and recall the rest, you see the rat's tail and conjure the image of the rat. Computational engines of that broad class prove extremely good at tasks such as sensori-motor co-ordination, face recognition, voice recognition, etc. But they are not well-suited to deductive logic, planning, and the typical tasks of sequential reason. They are, roughly speaking, "Good at Frisbee, Bad at Logic" – a cognitive profile that is at once familiar and alien. Familiar, because human intelligence clearly has something of that flavor. Yet alien, because we repeatedly transcend these limits, planning family vacations, running economies, solving complex sequential problems, etc., etc. A powerful hypothesis, which I first encountered in McClelland, Rumelhart, Smolensky and Hinton (1986), is that we transcend these limits, in large part, by combining the internal operation of a connectionist, pattern-completing device with a variety of external operations and tools which serve to reduce various complex, sequential problems to an ordered set of simpler pattern-completing operations of the kind our brains are most comfortable with. Thus, to borrow the classic illustration, we may tackle the problem of long multiplication by using pen, paper and numerical symbols. We then engage in a process of external symbol manipulations and storage so as to reduce the complex problem to a sequence of simple pattern-

completing steps that we already command, first multiplying 9 by 7 and storing the result on paper, then 9 by 6, and so on.The value of the use of pen, paper, and number symbols is thus that – in the words of Ed Hutchins;

> [Such tools] permit the [users] to do the tasks that need to be done while doing the kinds of things people are good at: recognizing patterns, modeling simple dynamics of the world, and manipulating objects in the environment.
>
> Hutchins (1995) p. 155

This description nicely captures what is best about *good* examples of cognitive technology: recent word-processing packages, web browsers, mouse and icon systems, etc. (It also suggests, of course, what is wrong with many of our first attempts at creating such tools – the skills needed to use those environments (early VCR's, word-processors, etc.) were *precisely* those that biological brains find hardest to support, such as the recall and execution of long, essentially arbitrary, sequences of operations. See Norman (1999) for discussion.

The conjecture, then, is that one large jump or discontinuity in human cognitive evolution involves the distinctive way human brains repeatedly create and exploit various species of cognitive technology so as to expand and re-shape the space of human reason. We – more than any other creature on the planet – deploy non-biological elements (instruments, media, notations) to *complement* our basic biological modes of processing, creating extended cognitive systems whose computational and problem-solving profiles are quire different from those of the naked brain.

The true significance of recent work on "embodied, embedded" problem-solving may thus lie not in the endless debates over the use or abuse of notions like internal representation, but in the careful depiction of complex, looping, multi-layered interactions between the brain, the body and reliable features of the local problem-solving environment. Internal representations will, almost certainly, feature in this story. But so will external representations, and artifacts, and problem-transforming tricks. The right way to "scale-up" the lessons of simple robotics so as to illuminate human thought and reason is to recognise

that human brains maintain an intricate cognitive dance with an ecologically novel, and immensely empowering, environment: the world of symbols, media, formalisms, texts, speech, instruments and culture. The computational circuitry of human cognition flows both within and beyond the head, through this extended network in ways which radically transform the space of human thought and reason. Modelling and understanding this environmentally *extended* cognitive device is an important task for any mechanistic account of human rationality.

4. Puzzles For Cyborgs.

Advanced human reason, according to the story just developed, is an *essentially extended* achievement. The nexus of resources that make our kind of rationality mechanically possible far outrun the resources of the individual biological brain, and include essential roles for non-biological props and artifacts, including external symbol systems. It's the props, tools and add-ons, according to this story (and the wider webs of encoding and processing in which they figure) that make us so special: that enable humans, more than any other animal on the planet, to glimpse and explore the spaces of reason.

This kind of story, with its general emphasis on the role of tools and equipment, clearly has roots both in Heidegger (1927/1961) and in the work of John Haugeland (e.g Haugeland (1998)). But it is probably most strongly associated, in the philosophical literature, with the work of Daniel Dennett (see especially Dennett (1995) (1996)).

One difference between Dennett's story and the kind of account suggested in section 3 above lies in their differing emphases on internalisation. The primary impact of the surrounding cognitive technologies lies, for Dennett, is the way exposure to them (especially to speech, which Dennett contentiously treats as a special kind of cognitive technology) "re-programs" the biological brain, allowing it to internalise routines and strategies first explored in some external arena. Thus we read that experience with the use of linguistic labels, inherited from our culture, sculpts individual brains in the ways distinctive of conceptualised understanding. Concepts (human-style concepts) are thus explained as 'artifactual modes in our memories...pale shadows of articulated

and heard words' (Dennett (1996) p.151). While I am not inclined to doubt that experience with words and labels adds a new dimension to human thought (see e.g. The discussion in Clark (1998)), I have tended (following Ed Hutchins- see Hutchins (1995)) to emphasise a somewhat different range of cases: cases in which processes of development and learning culminate *not* in the full internalisation of a (once) technologically-mediated competence, but in a kind of *dove-tailed distributed machine*. Thus recall the case of the painter whose creative activity involves a process of sketching and re-sketching. Or the architect whose design activity involves the use of CAD and virtual reality. Or the academic who carefully lays out her argument on paper for re-inspection and tweaking. Or the sailor who uses charts, alidades, nautical slide rules and the like (Hutchins (1995)). In all these cases - and the list is endless - the result of learning and development is not so much to internalise a once-external strategy, but to gear internal resources to play a specific role in a reliable, extended problem-solving matrix.

Now all this, I readily concede, is just a matter of emphasis. For something like full internalisation clearly *can* occur, and (contrariwise) Dennett also accepts the important role of what I have been calling "dovetailing". Thus Dennett even comments that:

> The primary source [of our greater intelligence]...is our habit of off-loading as much as possible of our cognitive tasks into the environment itself - extending our minds (that is, our mental projects and activities) into the surrounding world. (Dennett (1996) p.135).

Nonetheless, I think it is fair to say that, in general, Clark and Hutchins depict the primary role of cognitive technologies as expanding our effective cognitive architecture so as to include circuits and transformations that loop out into the surrounding environment, while Dennett has attended most closely to the developmental dimension in which external props sculpt and transform what is ultimately an internal resource.

The former image (of human cognitive architectures as literally extending outside the head) gives rise, however, to a number of questions and puzzles. A

common initial reaction is to ask How, if it is indeed all these designer environments that make us so smart, we were ever smart enough to build them in the first place? Such a question can be pressed in two quite different ways, requiring quite different kinds of response. The simplest formulation offers a (meetable) challenge: How can our designer environments be what makes us so smart, if we had to be that smart in order to build them in the first place? Thus formulated, there is a familiar (and correct) response: bootstrapping. On the bootstrapping scenario, you use the intelligence available to create a resource that makes *more* intelligence available, and repeat until satisfied (or extinct).

A better way to pose the question, however, is this: since no other species on the planet builds as varied, complex and open-ended designer environments as we do (the claim, after all, is that this is why we are *special*), what is it that allowed this process to get off the ground in our species in such a spectacular way? And isn't *that*, whatever it is, what really matters? Otherwise put, even if it's the designer environments that makes us so intelligent, what biological difference lets us build/discover/use them in the first place?

This is a serious, important and largely unresolved question. Clearly, there must be some (perhaps quite small) biological difference that lets us get our collective foot in the designer environment door - what can it be? The story I currently favour located the difference in a biological innovation for greater neural plasticity combined with the extended period of protected learning called "childhood" Thus Quartz (1999) and Quartz and Sejnowski (1997) present strong evidence for a vision of human cortex (especially the most evolutionarily recent structures such as neocortex and prefrontal cortex) as an "organ of plasticity" whose role is to dovetail the learner to encountered structures and regularities, and to allow the brain to make the most of reliable external problem-solving resources. This "neural constructivist" vision depicts neural (especially cortical) growth as experience - dependent, and as involving the actual construction of new neural circuitry (synapses, axons, dendrites) rather than just the fine-tuning of circuitry whose basic shape and form is already determined. One upshot is that the learning device *itself* changes as a result of organism-environmental interactions - learning does not just alter the knowledge base for a fixed computational engine, it alters the internal computational architecture itself. Evidence for this neural constructivist view comes primarily from recent

neuroscientific studies (especially work in developmental cognitive neuroscience). Key studies here include work involving cortical transplants, in which chunks of visual cortex were grafted into other cortical locations (such as somatosensory or auditory cortex) and proved plastic enough to develop the response characteristics appropriate to the new location (see Schlagger and O'Leary (1991)), work showing the deep dependence of specific cortical response characteristics on developmental interactions between parts of cortex and specific kinds of input signal (Chenn, (1997) and a growing body of constructivist work in Artificial Neural Networks: connectionist networks in which the architecture (number of units and layers, etc.) itself alters as learning progresses - see e.g. Quartz and Sejnowski (1997). The take home message is that immature cortex is surprisingly homogeneous, and that it 'requires afferent input, both intrinsically generated and environmentally determined, for its regional specialisation' (Quartz (1999) p.49).

So great, in fact, is the plasticity of immature cortex (and especially, according to Quartz and Sejnowski, that of prefrontal cortex) that O'Leary dubs it 'proto-cortex'. The linguistic and technological environment in which the brain grows and develops is thus poised to function as the anchor point around which such flexible neural resources adapt and fit. Such neural plasticity is, of course, not restricted to the human species (in fact, some of the early work on cortical transplants was performed on rats) , though our brains do look to be far and away the most plastic of them all. Combined with this plasticity, however, we benefit from a unique kind of developmental space- the unusually protracted human childhood.

In a recent evolutionary account which comports perfectly with the neural constructivist vision, Griffiths and Stotz (2000) argue that the long human childhood provides a unique window of opportunity in which "cultural scaffolding [can] change the dynamics of the cognitive system in a way that opens up new cognitive possibilities" (op cit p.11) These authors argue against what they nicely describe as the "dualist account of human biology and human culture" according to which biological evolution must first create the "anatomically modern human" and is then followed by the long and ongoing process of cultural evolution. Such a picture, they suggest, invites us to believe in something like a basic biological human nature, gradually co-opted and

obscured by the trappings and effects of culture and society. But this vision (which is perhaps not so far removed from that found in some of the more excessive versions of evolutionary psychology) is akin, they argue, to looking for the true nature of the ant by "removing the distorting influence of the nest" (op cit p.10). Instead we humans are, by nature, products of a complex and heterogenous developmental matrix in which culture, technology and biology are pretty well inextricably intermingled. The upshot, in their own words, is that:

> The individual representational system is part of a larger representational environment which extends far beyond the skin. Cognitive processes actually involve as components what are more traditionally conceived as the expressions of thought and the objects of thought. Situated cognition takes place within complex social structures which 'scaffold' the individual by means of artifactual, linguistic and institutional devices...[and]..culture makes humans as much as the reverse. (Griffiths and Stotz (2000) p.?).

In short it is a mistake to posit a biologically fixed "human nature" with a simple "wrap-around" of tools and culture. For the tools and culture are indeed as much determiners of our nature as products of it. Ours are (by nature) unusually plastic brains whose biologically proper functioning has always involved the recruitment and exploitation of non-biological props and scaffolds. More so than any other creature on the planet, we humans are *natural-born cyborgs*, factory tweaked and primed so as to participate in cognitive and computational architectures whose bounds far exceed those of skin and skull.

All this adds interesting complexity to recent evolutionary psychological accounts (see eg Pinker (1997)) which emphasize our ancestral environments. For we must now take into account a plastic evolutionary overlay which yields a constantly moving target, an extended cognitive architecture whose constancy lies mainly in its continual openness to change. For the more specific doctrine of biological cognitive incrementalism, the implications are even more unsettling. Even granting that the underlying biological innovations may have required just a small tweak to some ancestral repertoire, the upshot of this subtle alteration is

now a sudden, massive leap in cognitive-architectural space. For the cognitive machinery is now intrinsically geared to self-transformation, artifact-based expansion, and a snowballing/bootstrapping process of computational growth. The machinery of human reason (the environmentally extended apparatus of our distinctively human intelligence) turns out to be rooted in a biologically incremental progression while simultaneously existing on the far side of a precipitous cliff in cognitive-architectural space.

Other problems that may arise include fears concerning the fate, given some such story, of traditional conceptions of agency, responsibility and personhood. One problem, which can present itself as an attempted reductio of the whole 'extended mind' idea, concerns the threat of "mental bloat". The worry (discussed at greater length in Clark and Chalmers (1998)) is that allowing (to use our new example) the sketch-pad operations to count as part of the artist's own mental processes leads inevitably to absurdities such as counting the database of the dusty Encyclopedia Britannica, which I keep in my garage, as part of my general knowledge. Such intuitively pernicious extension is not, however, inevitable. It is quite proper, as we argue in Clark and Chalmers (1998), to restrict the props and aids which can count as part of *my* mental machinery to those that are, at the very least, reliably available when needed and used or accessed pretty much as automatically as biological processing and memory. But despite such concessions, the idea of the biological organism as the privileged locus of personhood, agency and responsibility dies hard. Thus Butler (1998) suggests that the ultimate locus of control and decision-making is *always* the biological brain. It might also be argued that extensions and dovetailing aside there is always a "core agent" identical with the biological self, that *conscious* thought must always supervene on only the biological aspects of any extended nexus, and that the biological machinery displays a kind of co-ordination and integration missing (or usually missing) in cases of organism-artifact problem-solving.

Let us grant, for the sake of argument, some kind of biological-individualistic restriction concerning the substrate of conscious thought. That costs us nothing, since no one, I suppose, identifies the agent or the person solely with the biological machinery that supports conscious thought. Yet all the other arguments, it seems to me, are prone to a single type of response viz: that reapplying each suggested criteria internally reveals it as both implausible and unprincipled. Let us ask, for example, whether my hippocampus participates in

ultimate decision-making? If it does not, are we to assume that the sub-personal hippocampal machinery is not part of the cognitive system that constitutes Andy-the-agent? Or suppose, as is not entirely out of the question, that there simply are no neural structures that reliably and constantly form the inner locus of "ultimate decision-making": that decisions emerge from the co-occurrence activity of many shifting coalitions of sub-systems. Once again, religious adherence to the suggested critirion would seem to lead to a kind of shrinkage in which the true cognitive agent is either identified with an unrealistically small sub-set of neural structures, or to an even more unnerving total disappearance, in which everything now appears external to some imaginary internal point source of genuine agency, control and will. Parallel arguments are easily constructed to show that strong notions of "core agency", "types of informational integration" and the like threaten, once re-applied inside the biological skin-bag, , to undermine the very notion (of the biological brain as the true engine of reason and the seat of personhood) they were supposed to protect.

Herbert Simon, in fact, was led to embrace just this kind of cognitive shrinkage, and for some quite revealing reasons. Simon saw, very clearly, that portions of the external world often functioned as a non-biological kind of memory. He thus saw the deep parity (parity, not identity) that can obtain between external and internal resources. But instead of counting those external (biological-organism-external) resources as proper parts of an extended knowing system, Simon chose to go the other way. Regarding biological, on-board memory, Simon chose to 'view this information-packed memory as less a part of the organism than of the environment to which it adapts' (Simon (1981) p.65). Part of the problem here no doubt originates from Simon's overly passive (mere storage) view of biological memory. But a deeper issue concerns the underlying image of something like a "core agent" surrounded by mere (internal and external) support systems (memories, etc.). It is this image which we must simply give up, if we are to preserve any substantial notion of the cognitive self. For the core, once we go searching for it, always turns out to be something thin and disappointing: in Simon's case, something not unlike the CPU of the traditional computer, or even the read-write head of the Turing machine!

A more promising approach, for the fans of biological individualism, might be to focus on providing a positive account of the value of a self-conception (an understanding of our own identity and personhood) which

continues to give pride of place to the contents of the ancient skin-bag. I doubt that such a positive account will be forthcoming. But the debate concerning biological/non-biological conceptions of the self is clearly far from over (it has, after all, been running ever since Locke (xxxx): for a nice overview, and a spirited defence of a non-biological conception, see Rovane (1998)).

Let us return, finally, to some questions closer to those with which we began: questions concerning the mechanical explanation of human reason itself. It is useful, with our previous discussion in mind, to notice that the kinds of reasoning and rationality philosophy has (at least in the present century) typically set out to explain and understand were, in fact, originally *situated*, and often socially distributed, achievements. Thus consider Ed Hutchins' revealing account of the kind of reasoning which the Turing machine model (see section 1 above) was originally supposed to capture. It was a model, Hutchins (and see Dennett (1991) chapter 7, Clark (1989) chapter 6) suggests, not ,in the first instance, of what goes on inside the individual head, but of the kinds of serial, symbol-matching problem-solving that we engage in using pen, paper and other external props and artifacts. The original inspiration was nothing other than the self-conscious process of breaking a problem into parts, applying rules, inscribing the results, and repeating the procedure again and again (see also Dennett (1991) p.212). In such cases:

> The mathematician or logician was visually and manually interacting with [symbols and artifacts in] the material world. [Since] the symbols are in the environment...the cognitive properties of the human are not the same as the cognitive properties of the system that is made up of the human in interaction with these symbols. (Hutchins (1995) p.361).

The image of serial, clunky, symbol manipulation as the deep explanation of the mechanical possibility of reason is thus plausibly taken as a repressed image of the *distributed mechanisms of environmentally situated reason*, rather than as a revealing image of the operation of the inner-biological aspects of that extended system. It is at precisely this point, I suggest, that the germs of truth in both the classicist and the connectionist visions of mechanical reason can be reconciled. For the connectionist has the better grip on the contribution of the

brain, while the classicist nicely captures certain features of some larger systems comprising brains, artifacts and external symbols. The lofty vantage point of situated reason provides the proper perch from which to view this old debate.

Further issues which bear signposting include the pressing need for an account of the features of an extended brain-artifact system which make it (hopefully) epistemologically sound: the features which allow such an extended organization to be truth-seeking and (hopefully) often truth-revealing. We here require an account of the distinctive roles, and interactive complexities, of the various (biological, social, artifactual) parts of such a distributed system, and of the conditions under which it can be understood as yielding knowledge. We need an account of what it is for a theory, idea or belief generated by such a distributed and heterogenous process to be reliable and/or justified. And we need (I suspect) an account personal responsibility and moral agency which respects the thin, de-centralized and distributed nature of situated intelligent control. None of this, I think, presents an intractable problem. But the required moral and cognitive epistemology (the epistemology of distributed and situated reason) does not yet exist. Its closest cousin is work in the philosophy and methodology of science. But overall, these are indeed philosophical projects for a new millenium.

Somewhat surprisingly, perhaps, Jerry Fodor (1994, lecture 4) lately shows himself not unsympathetic to a fairly close relative of this project. Fodor too is darkly suspicious of the vision of full cognitive incrementalism, commenting that 'only the most benighted of evolutionary gradualists could be sanguine that the apparently radical discontinuity between us and other creatures will prove to be merely quantitative' (op cit p.91). He thus embraces the project of discerning just what it is that is special about our own minds (op cit). And what is special, Fodor argues, is that we (alone) are typically aware of the contents of our own thoughts. But what matters most, it then seems, is not this initial skill so much as the subsequent snowball of designer-environment creation to which it gives rise. For such self-awareness, Fodor suggests, enables a creature:

> to construct, with malice aforethought, situations in which it will be caused to have the thought that P if and only if the thought is true. (Fodor (1994) p.92).

Awareness of the contents of our own thoughts, Fodor suggests, enables us to engage in repeated and cumulative exercises of cognitive self-management and, ultimately, ' to bootstrap our science' (op cit p.98). We design our worlds so that they cause us to believe more truths, which enables us to design newer and better worlds, which cause us to believe even *more* truths, and so on. This, for Fodor, is another key part of the story about how we manage to be 'largely rational creatures' (op cit p.102).[11]

Fodor thus offers an intriguingly different account of the underlying biological difference, but likewise stresses the power and potency of the subsequent cascade of designer environments.

Despite our many differences then, there is a surprising amount of common ground here emerging. For the most distinctive features of human reason, even Fodor seems willing to agree, depend not directly on the brute profile of the biological brain but on the larger, designer-environment involving, systems in which such brains now develop, operate and learn. A proper scientific model of human thought and reason thus requires serious, not merely peripheral, attention to the properties of the multiple larger scale structures in which human brains are so productively embedded: the spots where the biological rubber meets the purpose-built road.

4. Conclusions: Rationality at a Crossroads

The project of explaining how distinctively human kinds of thought and reason are mechanically possible is easily misconstrued. It is misconstrued as the project of understanding what is special about the human brain. No doubt there *is* something special about our brains. But understanding our peculiar profiles as reasoners, thinkers and knowers of our worlds requires an even broader perspective: one that targets multiple brains and bodies operating in specially constructed environments replete with artifacts, external symbols, and all the variegated scaffoldings of science, art and culture. Understanding what is

---

[11] Interestingly, Daniel Dennett has lately joined Fodor in this, stressing the importance of what he calls 'florid representing' - the self-aware use and deployment of representations, both internal and external. See Dennett (ms).

distinctive about human reason thus involves understanding the complementary contributions of both biology and (broadly speaking) technology, as well as the dense, reciprocal patterns of causal and co-evolutionary influence that run between them.

For us humans there is nothing quite so natural as to be bio-technological hybrids: cyborgs of an unassuming stripe. For we benefit from extended cognitive architectures comprising biological and non-biological elements, delicately intertwined. We are hybrids who occupy a region of design space radically different from those of our biological forbearers. It is for this reason that we should be wary of the apparently innocent and "naturalistic sounding" doctrine I have called "biological cognitive incrementalism".

Where, then, does this leave the reputedly fundamental question 'How is rationality mechanically possible?'. It leaves it, I think, at an important crossroads, uncertainly poised between the old and the new. For if the broad picture scouted above is even halfway correct, the full problem of explaining rationality becomes, precisely, the problem of explaining the production, in social, environmental, and technological context[12], of broadly appropriate adaptive response. Rationality (or as much of it as we humans typically enjoy) is what you get when this whole medley of factors are tuned and interanimated in a certain way. Figuring out this complex ecological balancing act just *is* figuring out how rationality is mechanically possible.

Andy Clark, August 2000

Bibliography

Ballard, D. (1991). "Animate Vision". *Artificial Intelligence* 48: 57-86

[12] Emotional context too, though I have not stressed that aspect here. For a nice treatment of the critical cognitive roles of emotion and feeling, see P. S. Churchland (1998).

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

Beer, R. (2000)

Brooks, R. (1991) "Intelligence without representation." *Artificial Intelligence* 47: 139-159

Butler, K. (1998). *Internal Affairs: A Critique of Externalism in the Philosophy of Mind*. Dordrecht, Kluwer.

Chambers, and Reisberg, (1989)

Chenn, A (1997) Development of the Cerebral Cortex in W. Cowan, T. Jessel and S. Ziputsky (eds) *Molecular and Cellular Approaches to Neural Development* Oxford, England, Oxford University Press 440-473

Churchland, P. M. (1989). *The Neurocomputational Perspective*. Cambridge, MIT/Bradford Books.

Churchland, P. M. (1995). *The Engine of Reason, the Seat of the Soul*. Cambridge, MA, MIT Press.

Churchland, P. S. (1998). "On The Contrary". *Feeling Reasons*. P. M. Churchland and P. S. Churchland, MIT Press: 231-254.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing*. Cambridge, MIT Press.

Clark, A. (1993). *Associative Engines: Connectionism, Concepts and Representational Change*. Cambridge, 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. (1998). Magic Words: How Language Augments Human Computation. *Language and Thought*. J. Boucher and P. Carruthers. Cambridge, Cambridge University Press.

Clark, A (1999) "An Embodied Cognitive Science?" *Trends In Cognitive Sciences* 3:9:1999: 345-351

Clark, A. (In press). *Mindware: An Introduction to the Philosophy of Cognitive Science.* Oxford University Press.

Clark, A. and Chalmers, D. (1998). "The Extended Mind." *Analysis* 58: 7-19.

Clark, A and Grush, R "Towards a Cognitive Robotics" *Adaptive Behavior* 7:1:1999, p. 5-16

Clark, A. and Toribio, J. (1994). "Doing Without Representing?" *Synthese*.

Dennett, D (1978) Why Not The Whole Iguana? *Behavioral and Brain Sciences* 1:103-4

Dennett, D. (1982). "Beyond Belief". *Thought and Object*. A. Woodfield. Oxford, Clarendon Press.

Dennett, D. (1991). *Consciousness Explained*. New York, Little Brown & Co.

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

Dennett, D. (1996). *Kinds of Minds*. New York, Basic Books.

Dennett, D. (1998). *Brainchildren: Essays on Designing Minds*. Cambridge, MA, MIT Press.

Dennett, D (ms)

Dreyfus, H. & Dreyfus, S. (1990). Making a mind versus modeling the brain: Artificial intelligence at a branch point. In M. Boden (ed) *The Philosophy of Artificial Intelligence* (p. 309-333). Oxford: Oxford University Press.

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

Fodor, J. (1994). *The Elm And The Expert.* Cambridge, MA, MIT Press.

Fodor, J. (1998) *In Critical Condition* Cambridge, MA, MIT Press

Gibson, J.J (1979) *The Ecological Approach To Visual Perception* Boston, MA Houghton-Mifflin

Godfrey-Smith, P. (1996). *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge University Press.

Griffiths, P. And Stotz, K. (ms). "How the Mind Grows".

Gigerenzer , G and Todd, P (1999) *Simple Heuristics That Make Us Smart* New York, Oxford University Press

Haugeland, J. (1981). Semantic Engines: An Introduction to Mind Design. In J. Haugeland (Ed), *Mind Design*. Cambridge, MA: MIT Press.

Haugeland, J. (1997). What is Mind Design? In J. Haugeland (Ed), *Mind Design II.* Cambridge, MA: MIT Press.

Haugeland, J (1998) Mind Embodied and Embedded, in J. Haugeland *Having Thought* Cambridge, MA, Mit Press

Hutchins , E (1995) *Cognition In The Wild* Cambridge, MA, Mit Press

Keijzer, F (1998). Doing Without Representations Which Specify What To Do *Philosophical Psychology* 11:3: 269-302

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

Kirsh, D and Maglio, P (1994) On Distinguishing Epistemic From Pragmatic Action
Cognitive Science 18:513-549

Lee and Reddish (1981)

Locke

Marr, D. (1981). "Artificial Intelligence: A Personal View". In J. Haugeland. *Mind Design.* Cambridge, MA, MIT Press. p. 129-142.

Marr, D. (1982) *Vision* San Francisco, Freeman

McClamrock,R (1995) *Existential Cognition* Chicago, University of Chicago Press

McClelland, Rumelhart, Smolensky and Hinton (1986)

Newell, A., & Simon, H. (1981). Computer Science as Empirical Enquiry. In J. Haugeland (ed), *Mind Design* . Cambridge: MIT Press.

Norman, D (1999) *The Invisible Computer* Cambridge, MA MIT Press

Pinker, S (1997) *How the Mind Works* New York, Norton

Quartz, S (1999) The Constructivist Brain *Trends In Cognitive Science* 3:2: 48-57

Quartz , S and Sejnowski, T (1997) The Neural Basis of Cognitive Development: A Constructivist Manifesto *Behavioral and Brain Sciences* 20:537-596

Rovane, C. (1998). *The Bounds of Agency*. Princeton University Press.

Schlagger, B and O'Leary, D (1991) Potential of Visual Cortex to Develop an Array of Functional Units Unique to Somatosensory Cortex *Science* 252 1556-1560

Simon, H (1981) *The Sciences of the Artificial* Cambridge, MA MIT Press

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

Stein, L (1994) Imagination and Situated Cognition *Journal of Experimental Artificial Intelligence* 6: 393-407

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

Tversky, A and Kahneman, D (1974) Judgements Under Uncertainty: Heuristics and Biases *Science* 185: 1124-1131

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

Van Leeuwen, C, Verstijnen,I and Hekkert, P (1999) Common unconscious dynamics underlie common conscious effects: a case study in the interactive nature of perception and creation. In S. Jordan (ed) *Modeling Consciousness Across the Disciplines* Lanhan, MD, University Press of America

Webb, B (1994) Robotic Experiments in Cricket Phonotaxis in D Cliff, P Husbands, J Meyer and S Wilson (eds)
From Animals to Animats 3 Cambridge Ma Mit Press 45-54

Webb, B (1996) A Cricket Robot *Scientific American* 275: 62-67