Mind & Language ISSN 0268-1064
Vol. 8 No. 4 Winter 1993

Article
Learning and Labeling

DANIEL C. DENNETT

Clark and Karmiloff-Smith (CKS) have written an extraordinarily valuable paper, which sympathetically addresses what has all too often been an acrimonious and ideology-ridden 'debate' and begins to transform it into a multi-perspective research program. By articulating the submerged hunches on both sides in a single framework, and adding some powerful new ideas of their own, they dispel much of the smoke of battle. What we can now see much more clearly is the need for a model of a brain/mind that, as they say, 'enriches itself from within by re-representing the knowledge that it has already represented'.

On the one hand, as they note, there is what I propose to call *ABC learning*. This is the foundational animal capacity to be gradually trained by an environment (in the wild or the laboratory), and thanks to the cumulative effect of several generations of theorists (Associationist to Behaviorist to Connectionist), we are getting quite clear about the strengths and limits of this real but not all-encompassing *variety* of learning. Although ABC learning can yield remarkably subtle and powerful discriminatory competences, capable of teasing out the patterns lurking in voluminous arrays of data, these competences tend to be anchored in the specific tissues that are modified by training. They are 'embedded' in the sense that they are incapable of being 'transported' readily to other data domains or other individuals. CKS note that while there are clear benefits to a design policy that 'intricately interweave[s] the various aspects of our knowledge about a domain in a single knowledge structure', there are costs as well: 'the interweaving makes it practically impossible to operate

Address for correspondence: Center for Cognitive Studies, 11 Miner Hall, Tufts University, Medford, MA 02155-7059, USA. Email: ddennett@pearl.tufts.edu.

on or otherwise exploit the various dimensions of our knowledge independently of one another'.[1]

So opaquely is such knowledge hidden in the mesh of the connections that, as CKS say, 'it is knowledge *in* the system, but it is not yet knowledge *to* the system'. Once we think of the contrast in these terms, it may remind us of many other 'intelligent' animal behaviors that are not trained but innate. For example, the unsettling and precocious singlemindedness with which the newly hatched cuckoo chick shoulders the competing eggs out of the nest in which it finds itself provokes what may be a comforting judgment: the evolutionary rationale for this behavior is crystal clear, but it is nothing *to* the cuckoo. The 'wisdom' of its behavior is in some sense embedded *in* the innate wiring that achieves this effect so robustly, but the cuckoo hasn't a clue about this rationale (see Dennett, 1987, 1991; Dretske, 1991; and Dennett, 1992). Why not? What would have to be added to the cuckoo's computational architecture for it to be able to appreciate, understand, exploit the wisdom interwoven in its neural nets?

A popular answer to this question, in its many guises, is 'symbols!' The answer is well nigh tautological, and hence is bound to be right, in *some* interpretation. How could it not be the case that implicit or tacit knowledge becomes explicit by being expressed or rendered in some medium of 'explicit' representation? Symbols, unlike the nodes woven into connectionist networks, are 'movable'; they can be 'manipulated'; they can be composed into larger structures where their contribution to the meaning of the whole can be a definite and generatable function of the structure—the syntactic structure—of the parts.

There is surely something right about this; what we human beings have that far outstrips the cognitive capacities of both rat and cuckoo (and *maybe* even outstrips the cognitive capacities of all other primates) is the capacity for swift, insightful learning—learning that does not depend on laborious training but is simply—'simply'!—ours as soon as we contemplate a suitable symbolic representation of the knowledge. (We do have to *understand* the representation we contemplate, of course, and there is where mystery still lurks.)

'So,' CKS ask, 'do we merely need to add to a connectionist network a mechanism that generates linguistic labels for the network's implicit knowledge?' How could we take a connectionist network and *turn it into* a symbolic system? Or better, for as the authors show, we need to leave the underlying connectionist power intact: How could we take a connectionist

---

[1] We must be careful not to exaggerate the embeddedness of interwoven knowlege. The trained pianist has subjected only his hands and fingers to laborious training in hitting the keys, but will probably find himself *already* more adept at playing the pedals of an organ with his feet than the untrained person. And as Karl Lashley famously proved years ago, what the rat learns in the maze is not merely a way of moving its legs till it gets to the food; it has learned something much more general: roughly, how to get to the food.

network and *attach* a symbolic system to it? (And if so, what would the symbolic system be 'made of' if not connectionist parts?) Or still better: How could we make a connectionist system *grow* a symbolic system on top of itself?

CKS are correct to identify me as one who has held that 'the human mind deploys essentially connectionist style representations but augments itself with the symbol structures of natural language in the public domain'. They go on to claim (p. 505):

> Theories that make the distinctive cognitive characteristics of humans dependent on an ability with public language seem, in general, to get the cart before the horse. It is more plausible to see our abilities with language as one *effect* or *product* of a deeper underlying difference in the redescriptive architecture of our cognitive apparatus, a difference that may group us with some non-linguistic higher mammals but separate us from hamsters, sea-slugs and standard connectionist networks.

They may be right, but I am inclined to think that they are overstating their case, and missing some of the very complexity they rightly insist upon. Does the advanced use of language depend on an RR capacity that develops, but is 'specified in innate predispositions', or does the RR capacity that develops depend to an important degree on what the child *acquires* when the pre-designed structures of natural language are moved from the child's enveloping culture into its brain? In what follows I want to defend, in an exploratory way, the idea that the capacity of a system to engage in representational redescription really does depend on that system's capacity—not yet fully developed, but in the process of development—to master and use a natural language. (As they say in their conclusion, the crucial question about the 'concrete computational mechanisms' that accomplish the transition to RR capacities remains unresolved, and I am making some sketchy suggestions about how it may come to be resolved.)

Karmiloff-Smith (1979) offered a pioneer expression of the now familiar idea that a main virtue of introducing higher-level representations is that one creates a new class of entities that can be operated upon, that can become 'objects of cognitive manipulation, transportable to other tasks' (CKS). But these very skills of cognitive manipulation have to be created along with the representations; that is, they have to develop out of something prior, some capacities that can be harnessed or exapted (to use Stephen Jay Gould's term) to the novel tasks of composing, saving, retrieving, revising, comparing these new internal objects. Karmiloff-Smith's own research with children gives us some of the best glimpses we have of children gradually equipping themselves with these competences, and I gladly concede that this process doesn't always directly involve the child's using natural language explicitly directed to the task in hand. I want to

suggest a few ways, however, in which what is going on during this process might nevertheless depend on the natural language competence the child is acquiring.

I begin with a useful analogy with a more recent technological breakthrough. The advent of high-speed still photography was a revolutionary technological advance for science because it permitted human beings, for the first time ever, to examine complicated temporal phenomena not in real time, but *in their own good time*—in leisurely, methodical, backtracking analysis of the traces they had created of those complicated events. Here a technological advance carried in its wake a huge enhancement in cognitive power. Before there were cameras and high speed film, there were plenty of observational and recording devices that permitted the scientist to extract data precisely from the world for subsequent analysis at his leisure. The exquisite diagrams and illustrations of several centuries of science are testimony to the power of these methods, but there is something special about a camera: it is 'stupid'. It does not have to understand its subject the way an artist or illustrator does in order to 'capture' the data represented in its products.

As I just noted, the sort of learning we human beings can achieve just by contemplating symbolic representations of knowledge depends not on our merely, in some sense, perceiving them, but also understanding them, and my rather curious suggestion is that in order to arrive at this marvellous summit, we must climb steps in which we *perceive but don't understand* our own representations.

Contemplating one's past experience in such a way as to make it good material for general judgments requires recording it, somehow, but recording one's past experience *in toto* is impossible. We are not equipped—though some like to think we are—with a sort of multi-media recording in the brain of all our experience. Recording 'edited' versions of our past experience would be possible if we had an initially 'stupid' way of doing both the editing and storing. (If we had to have a good understanding of what we were editing at the time we stored it, we would not need to take our time, later, to re-analyze and reconsider what we had done.) CKS discuss Mozer and Smolensky's promising idea of 'skeletonizing' networks, to extract the essential knowledge in them. They add the useful idea of skeletonizing copies of the networks, leaving the detailed, robust parent network intact for use in its original domain, and then, somehow, forming 'new structured representations' tied to these skeletonized copies. I am suggesting that the sophistications necessary to develop such a process are exapted from language-processes.

How might a *habit* of label-generation, hypothesis-formation and testing get started, and what is involved in the general practice of such 'redescription'? Nobody knows yet—certainly I don't know—but I have some speculations to offer that might not be too far wide of the mark. Consider what happens early in the linguistic life of any child. 'Hot!' says mother. 'Don't touch the stove!' At this point the child doesn't have to know what 'hot'

or 'touch' or 'stove' mean—they are *primarily* just sounds—auditory event-types that have a certain redolence, a certain familiarity, a certain echoing memorability to the child. They come to conjure up a situation-type, however, and not just a situation in which a specific prohibition is typically encountered but also a situation in which a certain auditory rehearsal is encountered.

We may crudely overstate the case and suppose that the child acquires the habit of saying to itself (aloud—why not?) 'Hot', 'Don't touch the stove!' without much of an idea what it means, as an associated part of the drill that goes with approaching and then avoiding the stove, but also as a sort of mantra that might be uttered at any other time. (*Cf.* Baddeley, 1984, *e.g.* on 'articulatory loops', a related idea, but rather differently positioned). After all, children are taken with the habit of rehearsing words they have just heard, in and out of context, building up recognition-links and association paths between the auditory properties and concurrent sensory properties, internal states, and so forth. That's a laughably crude sketch of the sort of process that must go on, but it could have the effect of initiating a habit of what we might call *semi-understood self-commentary*. The child, prompted initially by some insistent auditory associations provoked by its parents' admonitions, acquires the habit of adding a sound track to its activities, 'commenting' on them. The actual utterances would consist at the outset of large measures of 'scribble' (the nonsense-talk children engage in), real words mouthed with little or no appreciation of their meaning, and understood words. There would be mock exhortation, mock prohibition, mock praise, mock description, and all these would eventually mature into real exhortation, prohibition, praise and description. But the habit of adding the 'labels' would be driven into place before the labels had to be understood, even partially understood.

It is such initially 'stupid' practices, the mere 'mouthing' of labels in circumstances appropriate and inappropriate, I am suggesting, that could soon be turned into the habit of redescription. As the child lays down more associations between the auditory and articulatory processes, on the one hand, and other patterns of concurrent activity on the other, this would create 'nodes' of saliency in memory; a word can become *familiar* even without being understood. And it is these anchors of familiarity that could give a label an independent identity within the system. Without such independence, labels are 'invisible'.

Labeling is a non-trivial cognitive tactic, and it is worth a moment's digression to consider the conditions under which it works. Why does anyone ever label anything, and what does it take to label something? Suppose you were searching through thousands of boxes of shoes, looking for a housekey that you had good reason to believe had been hidden in one of them. Unless you are an idiot, or so frantic in your quest that you cannot pause to consider the wisest course, you will devise some handy scheme for cutting down your task by preventing you from looking more than once in each box. One way would be to move the boxes from one stack

(the unexamined stack) to another stack (the examined stack). Another way, potentially more energy efficient, is to put a check mark on each box as you examine it, and then adopt the rule never to bother looking in a box with a check mark on it. A check mark is a way of making the world simpler; it cuts down on your cognitive load by giving you a *simple* perceptual task in place of a more difficult—perhaps impossible—task. Notice that if the boxes are all lined up in a single row, and you don't have to worry about unnoticed re-orderings of the queue, you don't need to put check marks—you can just work your way from left to right, using the simple distinguisher nature has already provided you, the left/right distinction.

But now let's concentrate on the check mark itself. Will *anything* do as a checkmark? Clearly not. 'I put a faint smudge somewhere on each box as I examine it'. 'I bump the corner of each box as I examine it'. Not good choices, since the likelihood is too high that something else may already have inadvertently put such a mark on a box. I need something distinctive, something that I can be confident is the result of *my* labeling act, not some extraneously produced blemish. It should also be memorable, of course, so I will not be beset by confusions about whether or not this *is* my label, and if so, what policy I meant to be following when I adopted it. Only under these conditions will a label fulfil its *raison d'être*, which is to provide a cognitive crutch, off-loading a bit of cognitive work into the environment. This is perhaps the most primitive precursor of writing, the deliberate use of parts of the external world as 'peripheral' information-storage systems.

An interesting—and largely unasked, let alone unanswered—question is whether non-human animals ever engage in deliberate labeling or marking of this sort. There are the scent trails of insects and other animals, of course, and one can easily recognize their capacity to make various otherwise difficult cognitive tasks extremely easy. Many animals stake out territory by marking the boundary with urine or other idiosyncratic productions, but these are at least primarily for the information of other animals, not *aides-mémoire* for themselves. Clark's nuthatches are superbly good at locating the caches of seeds they have left behind, and they may use the debris they leave behind when they empty a cache as a sign to themselves that they needn't re-explore it (just like the shoe box check mark), but even if this is a good case (and I am tempted to think it is) it is a case of opportunistic exploitation of a disturbance that would be made in any case for other reasons. That is nature's way, of course, but the question is whether any other creatures—other than ourselves—have discovered the practice of *creating* labels for things for the express purpose of making their cognitive tasks easier.

Now return to the practice of *internal* labeling. The moral I want to draw from the digression about external labeling is that labels always need to be independently and readily identifiable, which means in this context that they must be ready *enhancers* of sought-for associations that are

already to some extent laid down in the system. Beyond that, they can be arbitrary, and their arbitrariness is actually part of what makes them distinctive—there is little risk of failing to notice the presence of the label—it doesn't just *blend into* its surroundings like a dent in the corner of a shoebox. It wears the deliberateness of its creation on its sleeve.

The habit of semi-understood self-commentary could, I am suggesting, be the origin of the practice of deliberate labeling, in words (or scribble-words or other private neologisms), which in turn could lead to a still more efficient practice, dropping all or most of the auditory and articulatory associations and just relying on the *rest* of the associations (and association-possibilities) to do the anchoring. The child, I suggest, can abandon such vehicles as out-loud mouthings, and create private, unvoiced neologisms as labels for features of its own activities.

We can take a linguistic object as a *found object* (even if we have somehow blundered into making it ourselves, rather than hearing it from someone else), and store it away for further consideration, 'off line'. This depends on there being a detachable guise for the label, something that is independent of meaning. Once we have created labels, and the habit of 'attaching' them to experienced circumstances, we have created a new class of objects that can themselves become the objects of all the pattern-recognition machinery, association-building machinery, and so forth. Like the scientists lingering retrospectively over an unhurried examination of the photographs they took in the heat of experimental battle, we can reflect, in recollection, on whatever patterns there are to be discerned in the various labeled exhibits we dredge out of memory.

As we improve, our labels become ever more refined, more perspicuous, ever better articulated, and the point is finally reached when we approximate (at least—and in fact at best) to the near magical prowess we began with: the *mere contemplation* of a representation is sufficient to call to mind all the appropriate lessons; we have become *understanders* of the objects we have created. We might call these artifactual nodes in our memories, these pale shadows of articulated and heard words, *concepts*. A concept, then, is an internal label which may or may not include among its many associations the auditory and articulatory features of a word (public or private). But words, I am suggesting, are the prototypes or forbears of concepts. The first concepts one can manipulate, I am suggesting, are 'voiced' concepts, and only concepts that can be manipulated can become objects of scrutiny for us.

Do animals have concepts? Does a dog have a concept of *cat*? Or *food*, or *master*? Yes and no. No matter how close extensionally a dog's 'concept' of cat is to yours, it differs radically in one way: the dog cannot *consider* its concept. It cannot ask itself if it knows what cats are; it cannot wonder whether cats are animals; it cannot attempt to distinguish the essence of cat (by its lights) from the mere accidents. Concepts are not things in the dog's world in the way cats are. Concepts are things in our world because we have language. No languageless mammal can have the concept of snow

the way we can, because such a mammal—a polar bear, let's say—has no way of *considering* snow 'in general' or 'in itself', and not for the trivial reason that it doesn't have a (natural language) *word* for snow, but because without a natural language, it has no talent for wresting concepts from their interwoven connectionist nests. There are good reasons for attributing to polar bears a *sort* of concept of snow. For instance, polar bears have an elaborate set of competences for dealing with snow in its various manifestations that are lacking in lions. We can speak of the polar bear's implicit or procedural knowledge of snow, and we can even investigate, empirically, the extension of the polar bear's *embedded* snow-concept, but then bear in mind that this is not a *wieldable* concept for the polar bear.[2]

I have expressed these hunches about the indirect dependence of RR on language in simple and extreme terms, for the sake of clarity. They are, or suggest, hypotheses that are empirically testable, and there is probably already plenty of evidence with which I am unfamiliar that points to major complications in my story, if it is to survive at all. It is possible, even likely, that a still more indirect (and, of course, convoluted) story would be closer to the truth, and that there are, as Clark and Karmiloff-Smith propose, important *innate* predispositions for internal labeling. But if there are, these might well be Baldwin-effect incorporations of Good Tricks that first were implemented in culturally transmitted linguistic (or 'proto-linguistic') habits (Dennett, 1991b, pp. 182–208). Seeing this link as a possibility might help us understand whatever we discover.

Center for Cognitive Studies
Tufts University
Medford, MA 02155
USA

References

Baddeley, A. 1984: Reading and Working Memory. *Visible Language*, 18, 311–22.
Dennett, D.C. 1987: *The Intentional Stance*. Cambridge, MA.: MIT Press.
Dennett, D.C. 1991a: Ways of Establishing Harmony. In B. McLaughlin (ed.), *Dretske and his Critics*. Oxford: Blackwell.
Dennett, D.C. 1991b: *Consciousness Explained*. Boston: Little Brown.
Dennett, D.C. 1992: La Compréhension Artisanale. In *Daniel C. Dennett et les Stratégies Intentionelles*, in *Lekton*, 2, Université de Québec à Montréal.
Dretske, F. 1991: Replies. In B. McLaughlin (ed.), *Dretske and his Critics*. Oxford: Blackwell.

[2] The discussion by Clark and Karmiloff-Smith of the ideas of Gareth Evans (1982) is relevant here–but there is no time on this occasion for me to discuss it.

Evans, G. 1982: *The Varieties of Reference*. Oxford: Oxford University Press.
Karmiloff-Smith, A. 1979: *A Functional Approach to Child Language*. Cambridge: Cambridge University Press.