This work originally appeared in:
Dennett, Daniel. 1994. Cognitive Science as Reverse Engineering: Several Meanings of "Top-Down" and "Bottom-Up". In *Logic, Methodology and Philosophy of Science IX*, ed. D. Prawitz, B. Skyrms and D. Westerstahl, p.p.680-689.

Logic, Methodology and Philosophy of Science IX
D. Prawitz, B. Skyrms and D. Westerstahl (Editors)

COGNITIVE SCIENCE AS REVERSE ENGINEERING
SEVERAL MEANINGS OF "TOP-DOWN" AND "BOTTOM-UP"

By DANIEL C. DENNETT
Center for Cognitive Studies, Tufts University, Medford, MA 02155

The vivid terms, "Top-down" and "Bottom-up" have become popular in several different contexts in cognitive science. My task today is to sort out some different meanings and comment on the relations between them, and their implications for cognitive science.

Models and methodologies

To a first approximation, the terms are used to characterize both research methodologies on the one hand, and models (or features of models) on the other. I shall be primarily concerned with the issues surrounding top-down versus bottom-up methodologies, but we risk confusion with the other meaning if we don't pause first to illustrate it, and thereby isolate it as a topic for another occasion. Let's briefly consider, then, the top-down versus bottom-up polarity in models of a particular cognitive capacity, language comprehension.

When a person perceives (and comprehends) speech, processes occur in the brain which must be partly determined bottom-up, by the input and partly determined top-down, by effects from on high, such as interpretive dispositions in the perceiver due to the perceiver's particular knowledge and interests. (Much the same contrast, which of course is redolent of Kantian themes, is made by the terms "data-driven" and "expectation-driven").

There is no controversy, so far as I know, about the need for this dual source of determination, but only about their relative importance, and when, where, and how the top-down influences are achieved. For instance, speech perception cannot be entirely data-driven because not only are the brains of those who know no Chinese not driven by Chinese speech in

the same ways as the brains of those who are native Chinese speakers, but also, those who know Chinese but are ignorant of, or bored by, chess-talk, have brains that will not respond to Chinese chess-talk in the way the brains of Chinese-speaking chess-mavens are. This is true even at the level of perception: what you hear and not just whether you notice ambiguities, and are susceptible to garden-path parsings, for instance is in some measure a function of what sorts of expectations you are equipped to have. Two anecdotes will make the issue vivid.

The philosopher Samuel Alexander, was hard of hearing in his old age, and used an ear trumpet. One day a colleague came up to him in the comnmon room at Manchester University, and attempted to introduce a visiting American philosopher to him. "THIS IS PROFESSOR JONES, FROM AMERICA!" he bellowed into the ear trumpet. "Yes, Yes, Jones, from America" echoed Alexander, smiling. "HE'S A PROFESSOR, OF BUSINESS ETHICS!" continued the colleague. "What?" replied Alexander. "BUSINESS ETHICS!" "What? Professor of what?" "PROFESSOR OF BUSINESS ETHICS!" Alexander shook his head and gave up: "Sorry. I can't get it. Sounds just like 'business ethics'!"

Alexander's comprehension machinery was apparently set with too strong a top-down component (though in fact he apparently perceived the stimulus just fine).

An AI speech-understanding system whose development was funded by DARPA (Defense Advanced Research Projects Agency), was being given its debut before the Pentagon brass at Carnegie Mellon University some years ago. To show off the capabilities of the system, it had been attached as the "front end" or "user interface" on a chess-playing program. The general was to play white, and it was explained to him that he should simply tell the computer what move lie wanted to make. The general stepped up to the mike and cleared his throat which the computer immmediately interpreted as "Pawn to King-4." Again, too much top-down, not enough bottom-up.

In these contexts, the trade-off between top-down and bottom-up is a design parameter of a model that might, in principle, be tuned to fit the circumstances. You might well want the computer to "hear" "Con to Ping-4" as "pawn to King-4" without even recognizing that it was making an improvement on the input. In these contexts, "top-down" refers to a contribution from "on high" from the central, topmost information stores to what is coming "up" from the transducers or sense organs. Enthusiasm for models that have provision for large top-down effects has waxed and waned over the years, from the euphoria of "new look" theories of perception, which emphasized the way perception went "beyond

the information given" in Jerry Bruner's oft-quoted phrase, to the dysphoria of Jerry Fodor's (1983) encapsulated modules, which are deemed to be entirely data-driven, utterly "cognitively impenetrable" to downward effects.

David Marr's (1982) theory of vision is a prime example of a model that stresses the power of purely bottom-up processes, which can, Marr stressed, squeeze a lot more out of the data than earlier theorists had supposed. The issue is complicated by the fact that the way in which Marr's model (and subsequent Marr-inspired models) squeeze so much out of the data is in part a matter of fixed or "innate" biases that amount to presuppositions of the machinery such as the so-called rigidity assumption that permits disambiguation of shape from motion under certain circumstances. Is the rigidity assumption tacitly embodied in the hardware a top-down contribution? If it were an optional hypothesis tendered for the nonce by the individual perceiver, it would be a paradigmatic top-down influence. But since it is a fixed design feature of the machinery, no actual transmission of "descending" effects occurs; the flow of information is all in one inward or upward direction. Leaving the further discussion of these matters for another occasion, we can use the example of Marr to highlight the difference between the two main senses of "top-down". While Marr, as I have just shown, was a champion of the power of bottom-up models of perception (at least in vision), he was also a main spokesperson for the top-down vision of methodology, in his celebrated three-level cascade of the computational, the algorithmic and the physical level. It is hopeless, Marr argued, to try to build cognitive science models from the bottom-up: by first modeling the action of neurons (or synapses or the molecular chemistry of neurotransmitter production), and then modeling the action of cell assemblies, and then tracts, and then whole systems (the visual cortex, the hippocampal system, the reticular system). You won't be able to see the woods for the trees. First, he insisted, you had to have a clear vision of what the task or function was that the neural machinery wvas designed to execute. This specification was at what he called, misleadingly, the computational level: it specified "the function" the machinery was supposed to compute and an assay of the inputs available for that computation. With the computational level specification in hand, he claimed, one could then make progress on the next level down, the algorithmic level, by specifying an algorithm (one of the many logically possible algorithms) that actually computed that function. Here the specification is constrained, somewhat, by the molar physical features of the machinery: maximum speed of computation, for instance, would restrict the class of algorithms, and so would macro-architectural features

dictating when and under what conditions various subcomponents could interact. Finally, with the algorithmic level more or less under control. one could address the question of actual implementation at the physical level.

Marr's obiter dicta on methodology gave compact and influential expression to what were already reigning assumptions in Artificial Intelligence. If AI is considered as primarily an engineering discipline, whose goal is to create intelligent robots or thinking machines, then it is quite obvious that standard engineering principles should guide the research activity: first you try to describe, as generally as possible, the capacities or competences you want to design, and then you try to specify, at an abstract level, how you would implement these capacities, and then, with these design parameters tentatively or defeasibly fixed, you proceed to the nitty-gritty of physical realization.

Certainly a great deal of research in AI probably the bulk of it is addressed to issues formulated in this top-down way. The sorts of questions addressed concern, for instance, the computation of three-dimensional structure from two-dimensional frames of input, the extraction of syntactic and semantic structure from symbol strings or acoustic signals, the use of meta-planning in the optimization of plans under various constraints, and so forth. The task to be accomplished is assumed (or carefully developed, and contrasted with alternative tasks or objectives) at the outset, and then constraints and problems in the execution of the task are identified and dealt with.

This methodology is a straightforward application of standard ("forward) engineering to the goal of creating artificial intelligences. This is how one designs and builds a clock, a water pump, or a bicycle, and so it is also how one should design and build a robot. The client or customer, if you like, describes the sought for object, and the client is the boss, who sets in motion a top-down process. This top-down design process is not simply a one-way street, however, with hierarchical delegation of unrevisable orders to subordinate teams of designers. It is understood that as subordinates attempt to solve the design problems they have been given, they are likely to find good reasons for recommending revisions in their own tasks, by uncovering heretofore unrecognized opportunities for savings, novel methods of simplifying or uniting subtasks, and the like. One expects the process to gravitate towards better and better designs, with not even the highest level of specification immune to revision. (The client said he wanted a solar-powered elevator, but has been persuaded, eventually, that a wind-powered escalator better fits his needs.)

Marr's top-down principles are an adaptation, then, of standard AI

methodology. Another expression of much the same set of attitudes is my distinction between the intentional stance, the design stance and the physical stance, and my characterization of the methodology of AI as the gradual elimination of the intentional through a cascade of homunculi. One starts with the ideal specification of an agent (a robot, for instance) in terms of what the agent ought to know or believe, and want, what information-gathering powers it should have, and what capacities for (intentional) action. It then becomes an engineering task to design such an intentional system, typically by breaking it up into organized teams of sub-agents, smaller, more stupid homunculi, until finally all the homunculi have been discharged replaced by machines. A third vision with the same inspiration is Allen Newell's distinction between what he calls the knowledge level and the physical symbol system level. It might seem at first that Newell simply lumps together the algorithmic level and the physical level, the design stance and the physical stance, but in fact he has made the same distinctions, while insisting, wisely, that it is very important for the designer to bear in mind the actual temporal and spatial constraints on architectures when working on the algorithmic level. So far as I can see, there is only a difference in emphasis between Marr, Newell and me on these matters.

What all three of us have had in common are several things:

(1) stress on being able (in principle) to specify the function computed (the knowledge level or intentional level) independently of the other levels.

(2) an optimistic assumption of a specific sort of functionalism: one that presupposes that the concept of the function of a particular cognitive system or subsystem can be specified. (It is the function which is to be optimally implemented.)

(3) A willingness to view psychology or cognitive science as reverse engineering in a rather straightforward way.

Reverse engineering is just what the term implies: the interpretation of an already existing artifact by an analysis of the design considerations that must have governed its creation.

There is a phenomenon analogous to convergent evolution in engineering: entirely independent design teams come up with virtually the same solution to a design problem. This is not surprising, and is even highly predictable, the more constraints there are, the better specified the task is. Ask five different design teams to design a wooden bridge to span a particular gorge and capable of bearing a particular maximum load, and it is to be expected that the independently conceived designs will be very similar: the efficient ways of exploiting the strengths and weaknesses of

wood are well-known and limited.

But when different engineering teams must design the same sort of thing a more usual tactic is to borrow from each other. When Raytheon wants to make an electronic widget to compete with General Electric's widget, they buy several of GE's widget, and proceed to analyze them: that's reverse engineering. They run them, benchmark them, x-ray them, take them apart, and subject every part of them to interpretive analysis: why did GE make these wires so heavy? What are these extra ROM registers for? Is this a double layer of insulation, and if so, why did they bother with it? Notice that the reigning assumption is that all these "why" questions have answers. Everything has a raison d'etre; GE did nothing in vain.

Of course if the wisdom of the reverse engineers includes a healthy helping of self-knowledge, they will recognize that this default assumption of optimality is too strong: sometimes engineers put stupid, pointless things in their designs, sometimes they forget to remove things that no longer have a function, sometimes they overlook retrospectively obvious shortcuts. But still, optimality must be the default asssumption; if the reverse engineers can't assume that there is a good rationale for the features they observe. they can't even begin their analysis.

What Marr, Newell, and I (along with just about everyone in AI) have long assumed is that this method of reverse engineering was the right way to do cognitive science. Whether you consider AI to be forward engineering (just build me a robot, however you want) or reverse engineering (prove, through building, that you have figured out how the human mechanism works), the same principles apply.

And within limits, the results have been not just satisfactory; they have been virtually definitive of cognitive science. That is, what makes a neuroscientist a cognitive neuroscientist, for instance, is the acceptance, to some degree, of this project of reverse engineering. One benefit of this attitude has been the reversal of a relentlessly stodgy and constructive attitude among some neuroscientists, who advocated abstention from all "speculation" that could not be anchored firmly to what is known about the specific activities in specific neural tracts with the result that they often had scant idea what they were looking for in the way of functional contrilution from their assemblies. (A blatant example would be theories of vision that could, with a certain lack of charity, be described as theories of television as if the task of the visual system were to produce an inner motion picture somewhere in the brain.)

But as Ramachandran (1985) and others (e.g., Hofstadter see Dennett, 1987) were soon to point out, Marr's top-down vision has its own blind spot: it over-idealizes the design problem. by presupposing first

that one could specify the function of vision (or of some other capacity of the brain), and second, that this function was optimally executed by the machinery.

That is not the way Mother Nature designs systems. In the evolutionary processes of natural selection, goal-specifications are not set in advance problems are not formulated and then proposed, and no selective forces guarantee optimal "solutions" in any case. If in retrospect we can identify a goal that has been optimally or suboptimally achieved by the evolutionary design process, this is something of a misrepresentation of history. This observation, often expressed by Richard Lewontin in his criticism of adaptationism, must be carefully put if it is to be anything but an attack on a straw man. Marr and others (including all but the silliest adaptationists) know perfectly well that the historical design process of evolution doesn't proceed by an exact analogue of the top-down engineering process, and in their interpretations of design they are not committing that simple fallacy of misimputing history. They have presupposed, however

and this is the target of a more interesting and defensible objection that in spite of the difference in the design processes, reverse engineering is just as applicable a methodology to systems designed by Nature, as to systems designed by engineers. Their presupposition, in other words, has been that even though the forward processes have been different, the products are of the same sort, so that the reverse process of functional analysis should work as well on both sorts of product.

A cautious version of this assumption would be to note that the judicious application of reverse engineering to artifacts already invokes the appreciation of historical accident. sub-optimal jury-rigs, and the like. so there is no reason why the same techniques, applied to organisms and their subsystems, shouldn't yield a sound understanding of their design. And literally thousands of examples of successful application of the techniques of reverse engineering to biology could be cited. Some would go so far (I am one of them) as to state that what biology is, is the reverse engineering of natural systems. That is what makes it the special science that it is and distinguishes it from the other physical sciences,

But if this is so, we must still take note of several further problems that make the reverse engineering of natural systems substantially more difficult than the reverse engineering of artifacts, unless we supplement it with a significantly different methodology. which might be called botton-up reverse engineering or, as its proponents prefer to call it: Artificial Life.

The Artificial Life movement (AL), inaugurated a few years ago with a conference at Los Alamos (Langton, 1989), exhibits the same early enthu-

siasm (and silly overenthusiasm) that accompanied the birth of AI in the early 60's. In my opinion, it promises to deliver even more insight than AI. The definitive difference between AI and AL is, I think, the role of bottom-up thinking in the latter. Let me explain.

A typical AL project explores the large scale and long range effects of the interaction between many small scale elements (perhaps all alike, perhaps populations of different types). One starts with a specification of the little bits, and tries to move towards a description of the behavior of the larger ensembles. Familiar instances that predate the official Artificial Life title are John Horton Conway's game of Life and other cellular automata, and, of course, connectionist models of networks, neural and otherwise. It is important to realize that connectionist models are just one family within the larger order of AL models.

One of the virtues of AL modeling strategies is a simple epistemic virtue: it is relatively easy to get interesting or surprising results. The neuroscientist Valentino Braitenberg, in his elegant little book, Vehicles: Experiments in Synthetic Psychology (1984), propounded what he called the law of uphill analysis and downhill synthesis, which states, very simply, that it much easier to deduce the behavioral competence of a system whose internal machinery you have synthesized than to deduce the internal machinery of a black box whose behavioral competence you have observed. But behind this simple epistemological point resides a more fundamental one. first noted, I think, by Langton.

When human engineers design something (forward engineering), they must guard against a notorious problem: unforeseen side effects. When two or more systems, well-designed in isolation, are put into a supersystem, this often produces interactions that were not only not part of the intended design, but positively harmful; the activity of one system inadvertently clobbers the activity of the other. By their very nature unforseeable by those whose gaze is perforce myopically restricted to the subsystem being designed, the only practical way to guard against unforeseen side effects is to design the subsystems to have relatively impenetrable boundaries that coincide with the epistemic boundaries of their creators. In short, you attempt to insulate the subsystems from each other, and insist on an overall design in which each subsystem has a single, well-defined function within the whole. The set of systems having this fundamental abstract architecture is vast and interesting, of course, but - and here is AL's most, persuasive theme it does not include very many of the systems designed by natural selection! The process of evolution is notoriously lacking in all foresight; having no foresight, unforeseen or unforeseeable side effects are nothing to it; it proceeds, unlike human engineers, via

the profligate process of creating vast numbers of relatively uninsulated designs, most of which, of course, are hopelessly flawed because of self-defeating side effects, but a few of which, by dumb luck, are spared that ignominious fate. Moreover, this apparently inefficient design philosophy carries a tremendous bonus that is relatively unavailable to the more efficient, top-down process of human engineers: thanks to its having no bias against unexamined side effects, it can take advantage of the very rare cases where beneficial serendipitous side effects emerge. Sometimes, that is, designs emerge in which systems interact to produce more than was aimed at. In particular (but not exclusively) one gets elements in such systems that have multiple functions.

Elements with multiple functions are not unknown to human engineering, of course, but their relative rarity is signaled by the delight we are apt to feel when we encounter a new one. One of my favorites is to be found in the Diconix portable printer: This optimally tiny printer runs on largish rechargeable batteries, which have to be stored somewhere: inside the platen or roller! On reflection. one can see that such instances of multiple function are epistemically accessible to engineers under various salubrious circumstances, but one can also see that by and large such solutions to design problems must be exceptions against a background of strict isolation of functional elements. In biology, one encounters quite crisp anatomical isolation of functions (the kidney is entirely distinct from the heart, nerves and blood vessels are separate conduits strung through the body), and without this readily discernible isolation, reverse engineering in biology would no doubt be humanly impossible, but one also sees superimposition of functions that apparently goes "all the way down" It is very. very hard to think about entities in which the elements have multiple overlapping roles in superimposed subsystems, and moreover, in which some of the most salient effects observable in the interaction of these elements may not be functions at all, but merely byproducts of the multiple functions being served.

If we think that biological systems and cognitive systems in particular are very likely to be composed of such multiple function, multiple effect, elements, we must admit the likelihood that top-down reverse engineering will simply fail to encounter the right designs in its search of design space. Artificial Life, then, promises to improve the epistemic position of researchers by opening up different regions of design space and these regions include the regions in which successful AI is itself apt to be found!

I will mention one likely instance. A standard feature of models of cognitive systems or thinkers or planners is the separation between a central

"workspace" or "working memory" and a long term memory. Materials are brought to the workspace to be considered, transformed, compared, incorporated into larger elements, etc. This creates what Newell has called the problem of "distal access" How does the central system reach out into the memory and find the right elements at the right time? This is reminiscent of Plato's lovely image of the aviary of knowledge, in which each fact is a bird, and the problem is to get the right bird to come when you call! So powerful is this image that most modelers are unaware of the prospect that there might be alternative images to consider and rule out. But nothing we know in functional neuroanatomy suggests anything like this division into separate workspace and nmemory. On the contrary, the sort of crude evidence we now have about activity in the cerebral cortex suggests that the very same tissues that are responsible for long term memory, thanks to relatively permanent adjustments of the connections, are also responsible, thanks to relatively fleeting relationships that are set up, for the transient representations that must be involved in perception and "thought". One possibility, of course, is that the two functions are just neatly superimposed in the same space like the batteries in the platen, but another possibility at least, an epistemic possibility it would be nice to explore is that this ubiquitous decomposition of function is itself a major mistake, and that the same effects can be achieved by machinery with entirely different joints. This is the sort of issue that can best be explored opportunistically the same way Mother Nature explores by bottom-up reverse engineering. To traditional top-down reverse engineering, this question is almost impervious to entry.

There are other issues in cognitive science that appear in a new guise when one considers the difference between top-down and bottom-up approaches to design, but a consideration of them is beyond the scope of this paper.

REFERENCES

1. BRAITENBERG, V., 1984, Vehicles: Experiments in Synthetic Psychology, Cambridge. MA: MIT Press/A Bradford Book.
2. DENNETT, D. C., 1987, "The Logical Geography of Computational Approaches: A View from the East Pole,", in M. Harnish and M. Brand, eds., Problems in the Representation of Knowledge, Tucson. AZ: University of Arizona Press.
3. FODOR, J. , 1983. The Modularity of Mind. Cambridge, MA: MIT Press/A Bradford Book.
4. LANGTON, C., 1989, Artificial Life. Redwood City. CA: Addison-Wesley.
5. MAHR, D., 1982 Vision, Sail Francisco: Freeman.

6. NEWELL, A.. YOST, G., LAIRD, J. E., ROSENBLOOM, P. S., and ALTMANN, 1990
"Formulating the Problem Space Computational Model, ", presented at the 25th Anniversary Symposium, School of Computer Science, Carnegie Mellon Univ. 24-26 September, 1990, forthcoming in R. Rashid, ed., ACM Pressbook, Reading, PA: Addison-Wesley.
7. RAMACHANDRAN, V. S., 1985 Guest Editorial in Perception, 14, pp.97 103.