[to appear in J. Khalfa, ed., *CAN INTELLIGENCE BE EXPLAINED?*, Oxford Univ. Press]

[In **1984** a workshop was held at the Maison Française in Oxford, organized by Jean Khalfa, and involving René Thom, Richard Gregory, myself and others. A volume was supposed to emerge from it, but it never appeared, so far as I know, and so this has never been published. (The collection of Darwin Lectures, *What is Intelligence?* (Cambridge Univ. Press, 1996) edited by Khalfa is an entirely different anthology.) I refer to this essay as 'forthcoming' in "Evolution, Error and Intentionality" in *The Intentional Stance*, but that was a promise never kept. -- DCD,February 15, 2008]

A route to intelligence: oversimplify and self-monitor

Daniel C. Dennett
Center for Cognitive Studies
Tufts University

I want to try to do something rather more speculative than the rest of you have done. I have been thinking recently about how one might explain some features of human reflective consciousness that seem to me to be very much in need of an explanation. I'm trying to see if these features could be understood as solutions to design problems, solutions arrived at by evolution, but also, in the individual, as a result of a process of unconscious self-design. I've been trying to think of this in the context of work in AI on the attempt to design intelligent robots – not "bed-ridden" expert systems, but systems that have to act in real time in the real world. If you want to think about something like this, you have to stray fairly far from experiments and hard empirical data; you have to get fairly speculative. Nevertheless the design efforts of people in AI do seem to bring home to conviction – if not to prove – various design constraints looming large and inescapable. If we can come to see why a system – or an organ or a behavior-pattern – must have certain features or a certain structure in order to do its task, this may help us ask the right questions, or at least keep us from dwelling on some of the wrong questions when we try to explain the machinery in the brain that is responsible for intelligent action.

Resuming the discussion of yesterday evening, let me remind you that intelligent action in the real world depends on anticipation, of two kinds: both the built-in, fast, unconscious modular anticipation of the sort we were considering yesterday, and, in the case of human beings and maybe some other higher species, something that looks much more like voluntary, conscious, expectation-formation and calculation about the future.

There is an important family of verbs that strangely enough has not yet been singled out for philosophical attention. Central members of the family are "avoid", "prevent", "hinder", "foster", and, perhaps the most basic of all, "change" in its transitive sense, where we think of one thing or agent or event "actively" changing something else. These are the pre-eminent verbs of *action*, where one is characterizing the situation in terms of a rational agent who, as one says, sets out the "change the course of history." This is a curious phrase. We all want to be able to change the course of history, if only in our own little corners of the world. The problem of free will is very much a matter of whether one thinks one can change the course of history, but of course this familiar phrase, on even the most superficial analysis, turns out to be deeply puzzling. If you suppose it is to be taken at its face value it is absurd. How could you change the course of history? From what to what? If history is simply the sequence of events that actually occur, then of course you can't change history. People say you can't change the past, and that's true enough, but then you can't change the future either.

When one is thinking in this mode in which one considers bringing about these changes that one so very much wants to bring about, one has to be thinking of an *anticipated* history, the way history is going to go *ceteris paribus*, the way history is going to go *unless* somebody does something, or *until* somebody does thing, or *in spite of* what somebody does. These verbs of agency can have no foothold outside the framework of a projected, anticipated history, even

when they are used to characterize the effects brought about by entirely inanimate objects. Let me illustrate this with an example borrowed from my book *Elbow Room* (Dennett, 1984). Imagine that astronomers discover a meteor heading for the earth, and they calculate that it is going to hit North America on Tuesday, and there is nothing anyone can do about it. People would be frantic, of course, wondering if there was anything to be done, and perhaps praying for miraculous deliverance from this terrible catastrophe. And then, suppose, on the eve of destruction, another meteor appears, plunging out of darkest space on a course that is just right to deflect the first onto a near-miss trajectory, thus narrowly *averting* the catastrophe, *preventing* calamity.

These words would come naturally to our lips on such an occasion. But am I suggesting that the second meteor was a miracle – a God-given answer to our prayers? No, I am supposing that the second meteor was always out there, tracing out exactly the intercepting course, just as predictably as the first; it was simply not noticed by the astronomers until the last minute. In fact, had they noticed the second meteor when they noticed the first, they would never have alarmed us, because (as they can see now in retrospect and could have calculated then) *there was never going to be a catastrophe*. It was merely an anticipated catastrophe – a mis-anticipated catastrophe. It seems appropriate to speak of an averted or prevented catastrophe because we compare an anticipated history with the way things turned out and we locate an event which was the "pivotal" event relative to the divergence between that anticipation and the actual course of events, and we call this the "act" of preventing or avoiding.

Mark Twain once said "I'm an old man, and I've seen many troubles, but most of them never happened." This is the experiential history of somebody who is used to living in the world of avoiding and preventing. This is the world in which a rational deliberator lives. Such a

deliberator has to have a world view that is constantly looking forward, anticipating the way things are going to go unless it does various things or until it does various things.

Suppose then that one wants to design a robot that will live in the real world and be capable of making decisions so that it can further its interests – whatever interests we artificially endow it with. We want in other words to design a foresightful planner. How must one structure the capacities – the representational and inferential or computational capacities – of such a being? The problem that such a creature faces is, as usual in Artificial Intelligence, the problem of combinatorial explosion. The way one obtains anticipations is by sampling the trajectories of things in one's perceptual world and using the information thus gathered to ground an inference or extrapolation about the future trajectory of the thing. One cannot deal intelligently with anything that one cannot track in this way. When I speak of tracking, I am not just thinking of tracking the trajectories through space of moving things, but also the trajectories through time of things like food stores, seasons, inflation rates, the relative political power of one's adversaries, one's credibility, and so forth. There are indefinitely many things that could be kept track of, but the attempt to track everything, to keep up-to-date information about everything, is guaranteed to lead to a self-defeating paroxysm of information-overload. No matter how much information one has about an issue, there is always more that one could have, and one can often know that there is more that one could have if only one were to take the time to gather it. There is always more deliberation possible, so the trick is to design the creature so that it makes reliable but not foolproof decisions within the deadlines naturally imposed by the events in its world that matter to it.

The fundamental problem, then, is what we might call the problem of Hamlet, who, you recall, frittered away his time in deliberation (or so it appears), vacillating and postponing. This

is the sort of postponement that Réné Thom was discussing yesterday in his example of the man at the crosswalk who must make a decision. One has to make decisions in real time, and this means that one has to do a less than perfect job if one is to succeed at all. So one must be designed from the outset to economize, to pass over *most* of the available information.

How then does one partition the task of the robot so that it is apt to make reliable real time decisions? One thing one can do is declare that some things in the world of the creature are to be considered *fixed*; no effort will be expended trying to track them, to gather more information on them. The state of these features is going to be set down in axioms, in effect, but these are built into the system *at no representational cost*. One simply designs the system in such a way that it works well provided the world is as one supposes it always will be, and makes no provision for the system to work well ("properly") under other conditions. The system as a whole operates *as if* the world were always going to be one way, so that whether the world really is that way is not an issue that can come up for determination. The rigid-linkage assumption in human vision described by Ullman, 1979, is a good example. It is presumably a design feature endorsed over the eons by natural selection. In the past, the important things that have moved in our visual neighborhoods have tended to be assemblages of linkages the parts of which are rigid (hands, wrists, arms, elbows, and so forth), and one can create a much more efficient visual system for a creature with such a world by simply building in the rigidity assumption. This permits very swift calculations for speedy identification and extrapolation of the futures of relevant parts of the world.

Other things in the world are to be declared as *beneath notice* even though they might in principle be noticeable were there any payoff to be gained thereby. These are things that are not fixed but the changes of which are of no direct relevance to the wellbeing of the creature. These

things are smeared into a blur, as it were, in our perceptual world and not further attended to. An example drawn from Wimsatt (1980) is the difference in cognitive strategy between two different predators: the insectivorous bird and the anteater, which both need to keep track of moving insects. The insectivorous bird tracks individual flying insects and samples their trajectories with a fast sampling technique: a very high flicker fusion rate relative to human vision. (If you showed a motion picture to such a bird, it would see it as a slide show, in effect, not continuous motion.) The bird sees the individual insects *as* individuals. The anteater does not track individual ants. The anteater sees swarms of ants as batches of edible substance. (If I believed it was always appropriate to speak this way, I would say that "ant" was a mass term in the anteater's language of thought!) It laps up regions of ant, and does not waste any of its cognitive resources tracking individual ants any more than we track individual molecules when we detect a "permeating" uniform odor in a volume of air which may contain a few parts per billion of the telltale molecule.

The "grain" of our own perception could be different; the resolution of detail is a function of our own calculus of wellbeing, given our needs and other capacities. In our design, as in the design of other creatures, there is a trade-off in the expenditure of cognitive effort and the development of effectors of various sorts. Thus the insectivorous bird has a trade-off between flicker fusion rate and the size of its bill. If it has a wider bill it can harvest from a larger volume in a single pass, and hence has a greater tolerance for error in calculating the location of its individual prey.

If then some of the things in the world are considered fixed, and others are considered beneath notice, and hence are just averaged over, this leaves the things that are changing and worth caring about. These things fall roughly into two divisions: the trackable and the chaotic.

The chaotic things are those things that we cannot routinely track, and for our deliberative purposes we must treat them as random, not in the quantum mechanical sense, and not even in the mathematical sense (e.g., as informationally incompresssible), but just in the sense of pseudo-random. These are features of the world which, given the expenditure of cognitive effort the creature is prepared to make, are untrackable; their future state is unpredictable.

This means that any real, finite deliberator must partition the states of its world in such a way as to introduce the concept of possibility: it is possible that item n is going to be in state A, and it is possible that item n is going to be in state B, or in state C. We get an ensemble of equipossible (but not necessarily equiprobable) alternatives. This idea of partitioning the world into "possible" alternatives that remain "open" is very clearly the introduction of a concept of *epistemic* possibility. It is what is possible relative to a particular agent's knowledge. As the agent gets more knowledge, this may contract the set of possibilities. "I used to think that state B was possible, but given what I just learned, I realize it is not possible." (See Dennett, 1984)

Sellars (1963, 1966) draws the very useful distinction between what he calls the manifest image and the scientific image. The manifest image is the everyday world view, the world of macroscopic, solid, colored objects, and other persons or rational agents. It is the world of folk physics and folk psychology. Then there is the scientific image: the world of atomic and sub-atomic particles too small to be perceived by the naked eye, the world of forces and light waves. Sellars draws his distinction in such a way as to focus on the manifest image shared by (normal) human beings, but I think we can usefully extend his distinction to other species. We are the only species that has developed science, and so we have a scientific image of the world, of the world that we and other species live in, in spite of the vast differences in our manifest images of that world. The manifest image enjoyed by a species is determined, I suggest, by the set of design

"decisions" that apportion things in its environment into the categories of fixed, or beneath notice, or trackable, or chaotic. (It is important to note that this way of thinking of the manifest image of a species somewhat belies the connotations of the adjective "manifest" – since it presupposes nothing about consciousness. It is not at all ruled out that an entirely unconscious creature – our imaginary robot, for instance – would have a manifest image.)

Why are we the only species to have developed a scientific image in addition to – and somewhat discordant with – our manifest image? That is a topic that has often been written on, so I will pause to make just one point. The principles of design that create a manifest image in the first place also create the loose ends that can lead to its unraveling. Some of the engineering shortcuts that are dictated if we are to avoid combinatorial explosion take the form of ignoring – treating as if non-existent – small changes in the world. They are analogous to "round off error" in computer number-crunching. And like round-off error, their locally harmless oversimplifications can accumulate under certain conditions to create large errors.

Then if the system can notice the large error, and diagnose it (at least roughly), it can begin to construct the scientific image. For example, we have been designed to detect "directly" only those changes that occur within a certain speed range. Outside our window of direct visibility lie those changes that happen too fast or too slow for us to perceive without the aid of time-lapse or slow-motion photography, for instance. We cannot see a plant or a child grow from moment to moment. We can see the sun's motion relative to the earth only at sunrise or sunset, or with the aid of a simple prosthetic extension of our senses – a couple of sticks stuck in the ground will do. But over a few minutes in the latter case, or months or years in the case of plants or children, we detect the difference: our expectations of no change (zero plus zero plus zero . . . equals zero) are overturned. Now the minimal, non-brilliant response to this is simply to make mid-course

corrections in our extrapolations of trajectory and continue as before. The insightful response is to notice that we have to do this (often) and to posit *changes too small to be seen*, the entering wedge into the scientific world of postulated, invisible phenomena. Thus it is from a variety of self-monitoring – in particular the noticing of a pattern in one's own cognitive responses – that the bounteous shift of vision arises.

Let me return to the manifest image of our foresighted planner, with its "open future" of types of epistemically possible events that matter to it but cannot normally be tracked by it. These are the alternatives it may deliberate about, and must deliberate about if it is to fend for itself in the world. One of the pre-eminent varieties of epistemically possible events is the category of the agent's own actions. These are systematically unpredictable by it. It can attempt to track and thereby render predictions about the decisions and actions of other agents, but (for fairly obvious and well-known logical reasons, familiar in the Halting Problem in computer science, for instance) it cannot make fine-grained predictions of its own actions, since it is threatened by infinite regress of self-monitoring and analysis. Notice that this does not mean that our creature cannot make some boundary-condition predictions of its own decisions and actions. Thus I can make reliable predictions about decisions I will make in the near future: tomorrow at breakfast I will decide how many cups of tea I will drink, and right now I predict that I will decide to have more than zero and less than four.

Now if our creature is to be able to choose among the alternatives of which it can conceive, what strategies of deliberation should we endow it with? One feature we want to build in is one mentioned by Réné Thom yesterday: we must guard against the possibility that an evaluation process will end in a tie – the classic problem of Buridan's ass. The cheap way of providing this safety measure is to build in something functionally analogous to a coin-flip: an

arbitrary, pseudo-random "oracle" available for a decision-aiding nudge whenever the system needs it. I am fascinated by Julian Jaynes' speculation (Jaynes 1976), that the various traditions of superstitious decision-making and prognostication found in the ancient world – throwing bones and lots, looking at the entrails of animals, consulting oracles, reading tea leaves – are actually stratagems more or less unconsciously invented by early human beings in order to get themselves out of the position of Buridan's ass, or out of the somewhat related predicament (Hamlet's, we might say) of one who simply does not know how to deliberate effectively about a complicated situation, and needs nevertheless to act somehow in a timely manner. When the issues are too imponderable, when one can think of no considerations that settle the issue, when one is simply at a loss as to how to continue deliberations, here, as in the case of the pedestrian at the crosswalk, it can be valuable simply to get yourself moving in one direction or another. It doesn't in the long run and on average matter which direction you move as long as you get out of your state of decisional funk and get a move on. These rituals, Jaynes suggests, had the effect of making up people's minds for them when they weren't very good at making up their own minds. So these were deliberative crutches, or prostheses. I mention them here because they provide a vivid example of something that was not designed and transmitted genetically by natural selection, but rather a cultural artifact, unconsciously designed by individuals.

To some ears the phrase "unconsciously designed" is an oxymoron, but what I mean is quite straightforward: by haphazard some individuals came to engage in these strange behaviors without having any point in mind, but they found they had agreeable results, and so under certain circumstances they became popular behaviors. And so the rituals were subjected to further design refinement and then preserved by cultural transmission. A behavioral strategy thus transmitted probably has no specific, organic (neural) control system (in computerese, no

"dedicated hardware"), but rather is just software, part of the "virtual machine" of the human decision-maker shaped by cultural and other environmental factors, and differently implemented in individual control structures.

The most fundamental problem that faces the designer of such a deliberator is what Artificial Intelligence calls the Frame Problem. Since I have described that unsolved problem at length elsewhere (Dennett, 1984a), I will just remark here that we may view it as the problem of the effective management of the manifest image of a planner, so that the sorts of informational or representational short-cuts taken yield anticipations that are both timely and reliable. It is called the Frame Problem because of the so-called frame axioms that apparently must be used to stipulate, systematically, the sorts of constancies of effect that are assumed in any particular manifest image. What are the (gross, reliable, normal) effects of moving one thing onto another, for instance? Can we codify this understanding into defining axioms for the action type *move x onto y*?

This should be a rather basic action in the repertoire of any interestingly capable agent, and will be immediately recognized by anyone who is familiar with the famous "blocks world" of AI – an imaginary table-top world consisting of a few colored, differently shaped blocks that can be moved around and stacked by an equally imaginary robot arm. (See SHRDLU, for instance, in Winograd 1972). This is a world of breathtaking simplicity compared to the real world of any even very simple creature. But even in this diminished world the frame problem looms large. Consider some of the frame axioms that are needed:

(1) If $z \neq x$, then if I move x onto y, then if z was on w before, z is on w afterwards.

(2) If z is blue, then if I move x onto y, z is blue afterwards.

(3) If z is red, then if I move x onto y, z is red afterwards.

Do we really need separate, independent axioms for everything that doesn't change? If we do, the definition of each action type is going to have to contain clauses for every predicate available for use in state descriptions in a mindless profusion of axioms – apparently an engineering monstrosity. Can we not have some more general, basic axioms, to the effect, for instance, that the colors of things don't change?

(4) (For all x) (If x is red, x stays red)

This won't do, since one of the action types we may want to include in the repertoire is *paint x red*, which rules out (4) and its kin on pain of contradiction. The unsolved problem is how to provide a system of world-knowledge representation that is both simple and efficient enough to avoid combinatorial explosion, while supple and sensitive enough to recover from at least some of the stupid effects of its deliberate oversimplification.

No one has a good solution to the Frame Problem yet, least of all me, but I would claim that one element in any good solution is going to be layers of self-noticing. I will close by describing briefly two examples of the sort of thing I have in mind. I once had a dog that loved to fetch tennis balls thrown to it, but faced with two balls on the lawn and unable to hold them both in his mouth at once, he would switch rapidly back and forth, letting go of one to grab the other, then seeing the dropped ball, and immediately emptying his mouth again to fetch it, and so forth. He would do this maybe twenty or thirty times, apparently acting on some oversimple rule to the effect that *getting* is better than *keeping*. This was a bad rule more or less built into him – he never unlearned it – but he didn't die of following it. That is, he wasn't so transfixed by the rule that he followed it until he dropped dead of starvation. Something would click over in him after those several dozen iterations and he would stop. He didn't have to know why he stopped.

He had a minimal safety valve – somehow sensitive to "excess" repetition of his own response – that stopped him, and let him set out on some more promising course of action.

A similar case was recently described by Geoffrey Hinton in a talk at MIT on the Boltzmann machine architecture he and Terry Sejnowski have developed (Hinton and Sejnowski, 1983, 1983a). Boltzmann machines are powerful problem solvers in certain traditionally difficult problem domains, but they have their characteristic weaknesses. Consider a typical problem graphically as the task of finding the lowest spot – the global minimum – in a large terrain dimpled with many depressions – local minima. (This is, of course, just "hill-climbing" turned upside down!) Boltzmann machines are efficient finders of global minima under many conditions, but they can be trapped in unusual terrains.

Consider a terrain crossed by a steep-sided gully, which slopes gently at the bottom towards the global minimum. When a Boltzmann machine "enters" such a gully in the course of its explorations, it asks itself, in effect, "which direction should I go to go down?" and looks around locally for the steepest downgrade. Only at the very bottom of the gully is the gentle slope towards the solution "visible"; at all other points the fall line (to use skier's jargon) will be at roughly right angles to that direction. With slight overshooting, the Boltzmann machine will end up somewhere on the opposite slope of the gully, ask its question again, and shoot back onto the opposite slope. Back and forth it will oscillate in the gully, oblivious to the futility of its search. Trapped in such an environment, a Boltzmann machine loses its normal speed and efficacy, and becomes a liability to any organism that relies on it.

As Hinton noted on the occasion, what one wants in such a situation is for the system to be capable of "noticing" that it had entered into such a repetitive cycle, and resetting itself on a different course. The design solution that thus might be favored is not to discard the Boltzmann

machine idea because it has this weakness, but to compensate for the weakness with some ad hoc strategy of oversight and management. Just this policy, I think, will be found to be endemic in the design of intelligent control systems.

Bibliography

Dennett, D. C., 1984, *Elbow Room: the Varieties of Free Will Worth Wanting*, Cambridge: Bradford Books/MIT Press, and Oxford: Oxford Univ. Press.

Dennett, D. C., 1984a, "Cognitive Wheels: the Frame Problem of AI," in C. Hookway, ed., *Minds, Machines and Evolution*, Cambridge Univ. Press. pp.129-151.

Hinton, G., and Sejnowski, J., 1983, "Optimal Perceptural Inference," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, Washington DC, June 1983.

Hinton, G., and Sejnowski, J., 1983a, "Analyzing Cooperative Computation," *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*, Rochester NY, May 1983.

Jaynes, J., 1976, *The Origins of Consciousness in the Breakdown of the Bicameral Mind*, Boston: Houghton Mifflin.

Sellars, W., 1963, *Science, Perception and Reality*, London: Routledge & Kegan Paul.

Sellars, W., 1966, "Fatalism and Determinism," in K. Lehrer, ed., *Freedom and Determinism*, New York: Random House.

Ullman, S., 1979, *The Interpretation of Visual Motion*, Cambridge: MIT Press.

Wimsatt, W., 1980, "Randomness and Perceived Randomness in Evolutionary biology," *Synthese*, 43, pp. 287-329.

Winograd, T., 1972, *Understanding Natural Language*, New York: Academic Press.