Cognitive wheels: the frame problem of AI

DANIEL C. DENNETT

(C. Hookway ed., *Minds, Machines and Evolution*, Cambridge University Press, 1984, pp. 129–150)

Once upon a time there was a robot, named R1 by its creators. Its only task was to fend for itself. One day its designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. R1 located the room, and the key to the door, and formulated a plan to rescue its battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT (Wagon, Room, t) would result in the battery being removed from the room. Straightaway it acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon. R1 knew that the bomb was on the wagon in the room, but didn't realize that pulling the wagon would bring the bomb out along with the battery. Poor R1 had missed that obvious implication of its planned act.

Back to the drawing board. 'The solution is obvious,' said the designers. 'Our next robot must be made to recognize not just the intended implications of its acts, but also the implications about their side-effects, by deducing these implications from the descriptions it uses in formulating its plans.' They called their next model, the robot-deducer, R1D1. They placed R1D1 in much the same predicament that R1 had succumbed to, and as it too hit upon the idea of PULLOUT (Wagon, Room, t) it began, as designed, to consider the implications of such a course of action. It had just finished deducing that pulling the wagon out of the room would not change the colour of the room's walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon—when the bomb exploded.

Back to the drawing board. 'We must teach it the difference between relevant implications and irrelevant implications,' said the designers, 'and teach it to ignore the irrelevant ones.' So they developed a method of tagging implications as either relevant or irrelevant to the project at hand, and installed the method in their next model, the robot-relevant-deducer, or R2D1 for short. When they subjected R2D1 to the test that had so unequivocally selected its ancestors for extinction, they were surprised to see it sitting, Hamlet-like, outside the room containing the ticking bomb, the native hue of its resolution sicklied o'er with the pale cast of thought, as Shakespeare (and more recently Fodor) has aptly put it. 'Do something!' they yelled at it. 'I am,' it retorted. 'I'm busily ignoring some thousands of implications I have determined to be irrelevant. Just as soon as I find an irrelevant implication, I put it on the list of those I must ignore, and...' the bomb went off.

All these robots suffer from the *frame problem*. If there is ever to be a robot with the fabled perspicacity and real-time adroitness of R2D2, robot-designers must solve the frame problem. It appears at first to be at best an annoying technical embarrassment in robotics, or merely a curious puzzle for the bemusement of people working in Artificial Intelligence (AI). I think, on the contrary, that it is a new, deep epistemological problem—accessible in principle but unnoticed by generations of philosophers—brought to light by the novel methods of AI, and still far from being solved. Many people in AI have come to have a similarly high regard for the seriousness of the frame problem. As one researcher has quipped, 'We have given up the goal of designing an intelligent robot, and turned to the task of designing a gun that will destroy any intelligent robot that anyone else designs!'

I will try here to present an elementary, non-technical, philosophical introduction to the frame problem, and show why it is so interesting. I have no solution to offer, or even any original suggestions for where a solution might lie. It is hard enough, I have discovered, just to say clearly what the frame problem is—and is not. In fact, there is less than perfect agreement in usage within the AI research community. McCarthy and Hayes, who coined the term, use it to refer to a particular, narrowly conceived problem about representation that arises only for certain strategies for dealing with a broader problem about real-time planning systems. Others call this broader problem the frame problem-'the whole pudding,' as Hayes has called it (personal correspondence)—and this may not be mere terminological sloppiness. If 'solutions' to the narrowly conceived problem have the effect of driving a (deeper) difficulty into some other quarter of the broad problem, we might better reserve the title for this hard-to-corner difficulty. With apologies to McCarthy and Hayes for

joining those who would appropriate their term, I am going to attempt an introduction to the whole pudding, calling it the frame problem. I will try in due course to describe the narrower version of the problem, 'the frame problem proper' if you like, and show something of its relation to the broader problem.

Since the frame problem, whatever it is, is certainly not solved yet (and may be, in its current guises, insoluble), the ideological foes of AI such as Hubert Dreyfus and John Searle are tempted to compose obituaries for the field, citing the frame problem as the cause of death. In What Computers Can't do (Dreyfus 1972), Dreyfus sought to show that AI was a fundamentally mistaken method for studying the mind, and in fact many of his somewhat impressionistic complaints about AI models and many of his declared insights into their intrinsic limitations can be seen to hover quite systematically in the neighbourhood of the frame problem. Dreyfus never explicitly mentions the frame problem, but is it perhaps the smoking pistol he was looking for but didn't quite know how to describe? Yes, I think AI can be seen to be holding a smoking pistol, but at least in its 'whole pudding' guise it is everyone's problem, not just a problem for AI, which, like the good guy in many a mystery story, should be credited with a discovery, not accused of a crime.

One does not have to hope for a robot-filled future to be worried by the frame problem. It apparently arises from some very widely held and innocuous-seeming assumptions about the nature of intelligence, the truth of the most undoctrinaire brand of physicalism, and the conviction that it must be possible to explain how we think. (The dualist evades the frame problem—but only because dualism draws the veil of mystery and obfuscation over all the tough how-questions; as we shall sec, thc problem arises when one takes seriously the task of answering certain how-questions. Dualists inexcusably excuse themselves from the frame problem.)

One utterly central—if not defining—feature of an intelligent being is that it can 'look before it leaps'. Better, it can think before it leaps. Intelligence is (at least partly) a matter of using well what you know—but for what? For improving the fidelity of your expectations about what is going to happen next, for planning, for considering courses of action, for framing further hypotheses with the aim of increasing the knowledge you will use in the future, so that you can preserve yourself, by letting your hypotheses die in your stead (as Sir Karl Popper once put it). The stupid—as opposed to ignorant—being is the one who lights the match to peer into the fuel tank, who saws off the limb he is sitting on, who locks his keys in his car and then spends the next hour wondering how on earth to get his family out of the car.

But when we think before we leap, how do we do it? The answer seems obvious: an intelligent being learns from experience, and then uses what it has learned to guide expectation in the future. Hume explained this in terms of habits of expectation, in effect. But how do the habits work? Hume had a hand-waving answer—associationism—to the effect that certain transition paths between ideas grew more likely-to-be-followed as they became well worn, but since it was not Hume's job, surely, to explain in more detail the mechanics of these links, problems about how such paths could be put to good use—and not just turned into an impenetrable maze of untraversable alternatives—were not discovered.

Hume, like virtually all other philosophers and 'mentalistic' psychologists, was unable to see the frame problem because he operated at what I call a purely semantic level, or a phenomenological level. At the phenomenological level, all the items in view are individuated by their meanings. Their meanings are, if you like, 'given'—but this just means that the theorist helps himself to all the meanings he wants. In this way the semantic relation between one item and the next is typically plain to see, and one just assumes that the items behave as items with those meanings ought to behave. We can bring this out by concocting a Humean account of a bit of learning.

Suppose that there are two children, both of whom initially tend to grab cookies from the jar without asking. One child is allowed to do this unmolested but the other is spanked each time she tries. What is the result? The second child learns not to go for the cookies. Why? Because she has had experience of cookie-reaching followed swiftly by spanking. What good does that do? Well, the idea of cookie-reaching becomes connected by a habit path to the idea of spanking, which in turn is connected to the idea of pain... so of course the child refrains. Why? Well, that's just the effect of that idea on that sort of circumstance. But why? Well, what else ought the idea of pain to do on such an occasion? Well, it might cause the child to pirouette on her left foot, or recite poetry, or blink, or recall her fifth birthday. But given what the idea of pain means, any of those effects would be absurd. True; now how can ideas be designed so that their effects are what they ought to be, given what they mean? Designing some internal things—an idea, let's call it—so that it behaves vis-a-vis its brethren as if it meant cookie or pain is the only way of endowing that thing with

that meaning; it couldn't mean a thing if it didn't have those internal behavioural dispositions.

That is the mechanical question the philosophers left to some dimly imagined future researcher. Such a division of labour might have been all right, but it is turning out that most of the truly difficult and deep puzzles of learning and intelligence get kicked downstairs by this move. It is rather as if philosophers were to proclaim themselves expert explainers of the methods of a stage magician, and then, when we ask them to explain how the magician does the sawing-the-lady-in-half trick, they explain that it is really quite obvious: the magician doesn't really saw her in half; he simply makes it appear that he does. 'But how does he do that ?' we ask. 'Not our department', say the philosophers—and some of them add, sonorously: 'Explanation has to stop somewhere.

When one operates at the purely phenomenological or semantic level, where does one get one's data, and how does theorizing proceed? The term 'phenomenology' has traditionally been associated with an introspective method—an examination of what is presented or given to consciousness. A person's phenomenology just was by definition the contents of his or her consciousness. Although this has been the ideology all along, it has never been the practice. Locke, for instance, may have thought his 'historical, plain method' was a method of unbiased self-observation, but in fact it was largely a matter of disguised aprioristic reasoning about what ideas and impressions had to be to do the jobs they 'obviously' did. The myth that each of us can observe our mental activities has prolonged the illusion that major progress could be made on the theory of thinking by simply reflecting carefully on our own cases. For some time now we have known better: we have conscious access to only the upper surface, as it were, of the multi-level system of information-processing that occurs in us. Nevertheless, the myth still claims its victims.

So the analogy of the stage magician is particularly apt. One is not likely to make much progress in figuring out how the tricks are done by simply sitting attentively in the audience and watching like a hawk. Too much is going on out of sight. Better to face the fact that one must either rummage around backstage or in the wings, hoping to disrupt the performance in telling ways; or, from one's armchair, think aprioristically about how the tricks must be done, given whatever is manifest about the constraints. The frame problem is then rather like the unsettling but familiar 'discovery' that so far as armchair thought can determine, a certain trick we have just observed is flat impossible.

Here is an example of the trick. Making a midnight snack. How is it that I can get myself a midnight snack? What could be simpler? I suspect there is some leftover sliced turkey and mayonniase in the fridge, and bread in the breadbox—and a bottle of beer in the fridge as well. I realize I can put these elements together, so I concoct a childishly simple plan: I'll just go and check out the fridge, get out the requisite materials, and make myself a sandwich, to be washed down with a beer. I'll need a knife, a plate, and a glass for the beer. I forthwith put the plan into action and it works! Big deal.

Now of course I couldn't do this without knowing a good deal—about bread, spreading mayonniase, opening the fridge, the friction and inertia that will keep the turkey between the bread slices and the bread on the late as I carry the plate over to the table beside my easy chair. I also need to know about how to get the beer out of the bottle into the glass. Thanks to my previous accumulation of experience in the world, fortunately, I am equipped with all this worldly knowledge. Of course some of the knowledge I need might be innate. For instance, one trivial thing I have to know is that when the beer gets into the glass it is no longer in the bottle, and that if I'm holding the mayonnaise jar in my left hand I cannot also be spreading the mayonnaise with the knife in my left hand. Perhaps these are straightforward implications - instantiations—of some more fundamental things that I was in effect born knowing such as, perhaps, the fact that if something is in one location it isn't also in another, different location; or the fact that two things can't be in the same place at the same time; or the fact that situations change as the result of actions. It is hard to imagine just how one could learn these facts from experience.

Such utterly banal facts escape our notice as we act and plan, and it is not surprising that philosophers, thinking phenomenologically but introspectively, should have overlooked them. But if one turns one's back on introspection, and just thinks 'hetero-phenomenologically' about the purely informational demands of the task—what must be known by any entity that can perform this task—these banal bits of knowledge rise to our attention. We can easily satisfy ourselves that no agent that did not in some ways have the benefit of the information (that beer in the bottle is not in the glass, etc.) could perform such a simple task. It is one of the chief methodological beauties of AI that it makes one be a phenomenologist in this improved way. As a hetero-phenomenologist, one reasons about what the agent must 'know' or figure out unconsciously or

consciously in order to perform in various ways.

The reason AI forces the banal information to the surface is that the tasks set by AI start at zero: the computer to be programmed to simulate the agent (or the brain of the robot, if we are actually going to operate in the real, non-simulated world), initially knows nothing at all 'about the world'. The computer is the fabled tabula rasa on which every required item must somehow be impressed, either by the programmer at the outset or via subsequent 'learning' by the system.

We can all agree, today, that there could be no learning at all by an entity that faced the world at birth as a tabula rasa , but the dividing line between what is innate and what develops maturationally and what is actually learned is of less theoretical importance than one might have thought. While some information has to be innate, there is hardly any particular item that must be: an appreciation of modus ponens, perhaps, and the law of the excluded middle, and some sense of causality. And while some things we know must be learned—e.g. that Thanksgiving falls on a Thursday, or that refrigerators keep food fresh-many other 'very empirical' things could in principle be innately known—e.g. that smiles mean happiness, or that unsuspended, unsupported things fall. (There is some evidence, in fact, that there is an innate bias in favour of perceiving things to fall with gravitational acceleration.)

Taking advantage of this advance in theoretical understanding (if that is what it is), people in AI can frankly ignore the problem of learning (it seems) and take the shortcut of installing all that an agent has to 'know' to solve a problem. After all, if God made Adam as an adult who could presumably solve the midnight snack problem ab initio, AI agent-creators can in principle make an 'adult' agent who is equipped with worldly knowledge as if it had laboriously learned all the things it needs to know. This may of course be a dangerous short cut.

The installation problem is then the problem of installing in one way or another all the information needed by an agent to plan in a changing world. It is a difficult problem because the information must be installed in a usable format. The problem can be broken down initially into the semantic problem and the syntactic problem. The semantic problem called by Allen Newell the problem at the 'knowledge level' (Newell 1982)—is the problem of just what information (on what topics, to what effect) must be installed. The syntactic problem is what system, format, structure, or mechanism to use to put that information in.

The division is clearly seen in the example of the midnight snack problem. I listed a few of the very many humdrum facts one needs to know to solve the snack problem, but I didn't mean to suggest that those facts are stored in me—or in any agent—piecemeal, in the form of a long list of sentences explicitly declaring each of these facts for the benefit of the agent. That is of course one possibility, officially: it is a preposterously extreme version of the 'language of thought' theory of mental representation, with each distinguishable 'proposition' separately inscribed in the system. No one subscribes to such a view; even an encyclopedia achieves important economics of explicit expression via its organization, and a walking encyclopedia - not a bad caricature of the envisaged AI agent—must use different systemic principles to achieve efficient representation and access. We know trillions of things; we know that mayonnaise doesn't dissolve knives on contact, that a slice of bread is smaller than Mount Everest, that opening the refrigerator doesn't cause a nuclear holocaust in the kitchen.

There must be in us—and in any intelligent agent—some highly efficient, partly generative or productive system of representing—storing for use—all the information needed. Somehow, then, we must store many 'facts' at once—where facts are presumed to line up more or less one-to-one with non-synonymous declarative sentences. Moreover, we cannot realistically hope for what one might call a Spinozistic solution - a small set of axioms and definitions from which all the rest of our knowledge is deducible on demand—since it is clear that there simply are no entailment relations between vast numbers of these facts. (When we rely, as we must, on experience to tell us how the world is, experience tells us things that do not at all follow from what we have heretofore known.)

The demand for an efficient system of information storage is in part a space limitation, since our brains are not all that large, but more importantly it is a time limitation, for stored information that is not reliably accessible for use in the short real-time spans typically available to agents in the world is of no use at all. A creature that can solve any problem given enough time—say a million years—is not in fact intelligent at all. We live in a time-pressured world and must be able to think quickly before we leap. (One doesn't have to view this as an a priori condition on intelligence. One can simply note that we do in fact think quickly, so there is an empirical question about how we manage to do it.)

The task facing the AI researcher appears to be designing a system that can plan by using well-selected elements from its store of knowledge about the world it operates in. 'Introspection' on how we plan yields the following description of a process: one envisages a certain situation (often very sketchily); one then imagines performing a certain act in that situation; one then 'sees' what the likely outcome of that envisaged act in that situation would be, and evaluates it. What happens backstage, as it were, to permit this 'seeing' (and render it as reliable as it is) is utterly inaccessible to introspection.

On relatively rare occasions we all experience such bouts of thought, unfolding in consciousness at the deliberate speed of pondering. These are occasions in which we are faced with some novel and relatively difficult problem, such as: How can I get the piano upstairs? or Is there any way to electrify the chandelier without cutting through the plaster ceiling? It would be quite odd to find that one had to think that way (consciously and slowly) in order to solve the midnight snack problem. But the suggestion is that even the trivial problems of planning and bodily guidance that are beneath our notice (though in some sense we 'face' them) are solved by similar processes. Why? I don't observe myself planning in such situations. This fact suffices to convince the traditional, introspective phenomenologist that no such planning is going on. The hetero-phenomenologist, on the other hand, reasons that one way or another information about the objects in the situation, and about the intended effects and side-effects of the candidate actions, must be used (considered, attended to, applied, appreciated). Why? Because otherwise the 'smart' behaviour would be sheer luck or magic. (Do we have any model for how such unconscious information-appreciation might be accomplished? The only model we have so far is conscious, deliberate information-appreciation. Perhaps, AI suggests, this is a good model. If it isn't, we are all utterly in the dark for the time being.)

We assure ourselves of the intelligence of an agent by considering counterfactuals: if I had been told that the turkey was poisoned, or the beer explosive, or the plate dirty, or the knife too fragile to spread mayonnaise, would I have acted as I did? If I were a stupid 'automaton'—or like the Sphex wasp who 'mindlessly' repeats her stereotyped burrow-checking routine till she drops—I might infelicitously 'go through the motions' of making a midnight snack oblivious to the recalcitrant features of the environment.' But in fact, my midnight-snack-making behaviour is multifariously sensitive to current and background information about the situation. The only way it could be so sensitive—runs the tacit hetero-phenomenological reasoning—is for it to examine, or test for, the information in question. The information manipulation may be unconscious and swift, and it need not (it better not) consist of hundreds or thousands of seriatim testing procedures, but it must occur somehow, and its benefits must appear in time to help me as I commit myself to action.

I may of course have a midnight snack routine, developed over the years, in which case I can partly rely on it to pilot my actions. Such a complicated 'habit' would have to be under the control of a mechanism of some complexity, since even a rigid sequence of steps would involve periodic testing to ensure that subgoals had been satisfied. And even if I am an infrequent snacker, I no doubt have routines for mayonnaise-spreading, sandwich-making, and getting-something-out-of-the-fridge, from which I could compose my somewhat novel activity. Would such ensembles of routines, nicely integrated, suffice to solve the frame problem for me, at least in my more 'mindless' endeavours? That is an open question to which I will return below.

It is important in any case to acknowledge at the outset, and remind oneself frequently, that even very intelligent people do make mistakes; we are not only not infallible planners; we are quite prone to overlooking large and retrospectively obvious flaws in our plans. This foible manifests itself in the familiar case of 'force of habit' errors (in which our stereotypical routines reveal themselves to be surprisingly insensitive to some portentous environmental changes while surprisingly sensitive to others). The same weakness also appears on occasion in cases where we have consciously deliberated with some care. How often have you embarked on a project of the piano-moving variety—in which you've thought through or even 'walked through' the whole operation in advance—only to discover that you must backtrack or abandon the project when some perfectly foreseeable but unforeseen obstacle or unintended side-effect loomed? If we smart folk seldom actually paint ourselves into corners, it may be not because we plan ahead so well as that we supplement our sloppy planning powers with a combination of recollected lore (about fools who paint themselves into corners, for instance) and frequent progress checks as we proceed. Even so, we must know enough to call up the right lore at the right time, and to recognize impending problems as such.

To summarise: we have been led by fairly obvious and compelling considerations to the conclusion that an intelligent agent must engage in swift information-sensitive 'planning' which has the effect of producing

reliable but not foolproof expectations of the effects of its actions. That these expectations are normally in force in intelligent creatures is testified to by the startled reaction they exhibit when their expectations are thwarted. This suggests a graphic way of characterizing the minimal goal that can spawn the frame problem: we want a midnight-snack-making robot to be 'surprised' by the trick plate, the unspreadable concrete mayonnaise, the fact that we've glued the beer glass to the shelf. To be surprised you have to have expected something else, and in order to have expected the right something else, you have to have and use a lot of information about the things in the world.

The central role of expectation has led some to conclude that the frame problem is not a new problem at all, and has nothing particularly to do with planning actions. It is, they think, simply the problem of having good expectations about any future events, whether they are one's own actions, the actions of another agent, or the mere happenings of nature. That is the problem of induction—noted by Hume and intensified by Goodman (Goodman 1965), but still not solved to anyone's satisfaction. We know today that the problem of induction is a nasty one indeed. Theories of subjective probability and belief fixation have not stabilized in reflective equilibrium, so it is fair to say that no one has a good, principled answer to the general question: given that I believe all this (have all this evidence), what ought I to believe as well (about the future, or about unexamined parts of the world)?

The reduction of one unsolved problem to another is some sort of progress, unsatisfying though it may be, but it is not an option in this case. The frame problem is not the problem of induction in disguise. For suppose the problem of induction were solved. Suppose—perhaps miraculously - that our agent has solved all its induction problems or had them solved by fiat; it believes, then, all the right generalizations from its evidence, and associates with all of them the appropriate probabilities and conditional probabilities. This agent, ex hypothesi, believes just what it ought to believe about all empirical matters in its ken, including the probabilities of future events. It might still have a bad case of the frame problem, for that problem concerns how to represent (so it can be used) all that hard-won empirical information - a problem that arises independently of the truth value, probability, warranted assertability, or subjective certainty of any of it. Even if you have excellent knowledge (and not mere belief) about the changing world, how can this knowledge be represented so that it can be efficaciously brought to bear?

Recall poor R1D1 and suppose for the sake of argument that it had perfect empirical knowledge of the probabilities of all the effects of all its actions that would be detectable by it. Thus it believes that with probability 0.7864, executing PULLOUT (Wagon, Room) will cause the wagon wheels to make an audible noise; and with probability 0.5, the door to the room will open in rather than out; and with probability 0.999996, there will be no live elephants in the room, and with probability 0.997 the bomb will remain on the wagon when it is moved. How is R1D1 to find this last, relevant needle in its haystack of empirical knowledge? A walking encyclopedia will walk over a cliff, for all its knowledge of cliffs and the effects of gravity, unless it is designed in such a fashion that it can find the right bits of knowledge at the right times, so it can plan its engagements with the real world.

The earliest work on planning systems in AI took a deductive approach. Inspired by the development of Robinson's methods of resolution theorem proving, designers hoped to represent all the system's 'world knowledge' explicitly as axioms, and use ordinary logic—the predicate calculus - to deduce the effects of actions. Envisaging a certain situation S was modelled by having the system entertain a set of axioms describing the situation. Added to this were background axioms (the so-called 'frame axioms' that give the frame problem its name) which describe general conditions and the general effects of every action type defined for the system. To this set of axioms the system would apply an action - by postulating the occurrence of some action A in situation S - and then deduce the effect of A in S, producing a description of the outcome situation S'. While all this logical deduction looks like nothing at all in our conscious experience, research on the deductive approach could proceed on either or both of two enabling assumptions: the methodological assumption that psychological realism was a gratuitous bonus, not a goal, of 'pure' AI, or the substantive (if still vague) assumption that the deductive processes described would somehow model the backstage processes beyond conscious access. In other words, either we don't do our thinking deductively in the predicate calculus but a robot might; or we do (unconsciously) think deductively in the predicate calculus. Quite aside from doubts about its psychological realism, however, the deductive approach has not been made to work—the proof of the pudding for any robot—except for deliberately trivialized cases.

Consider some typical frame axioms associated with the action-type: move x onto y.

(1) If z ? x and I move x onto y, then if z was on w before, then z is on w after.

(2) If x is blue before, and I move x onto y, then x is blue after.

Note that (2), about being blue, is just one example of the many boring 'no-change' axioms we have to associate with this action-type. Worse still, note that a cousin of (2), also about being blue, would have to be associated with every other action-type—with pick up x and with give x to y, for instance. One cannot save this mindless repetition by postulating once and for all something like

(3) If anything is blue, it stays blue,

for that is false, and in particular we will want to leave room for the introduction of such action-types as paint x red. Since virtually any aspect of a situation can change under some circumstance, this method requires introducing for each aspect (each predication in the description of S) an axiom to handle whether that aspect changes for each action-type.

This representational profligacy quickly gets out of hand, but for some 'toy' problems in AI, the frame problem can be overpowered to some extent by a mixture of the toyness of the environment and brute force. The early version of SHAKEY, the robot at SRI, operated in such a simplified and sterile world, with so few aspects it could worry about that it could get away with an exhaustive consideration of frame axioms.

Attempts to circumvent this explosion of axioms began with the proposal that the system operate on the tacit assumption that nothing changes in a situation but what is explicitly asserted to change in the definition of the applied action (Fikes and Nilsson 1971). The problem here is that, as Garrett Hardin once noted, you don't do just one thing. This was R1's problem, when it failed to notice that it would pull the bomb out with the wagon. In the explicit representation (a few pages back) of my midnight snack solution, I mentioned carrying the plate over to the table. On this proposal, my model of S' would leave the turkey back in the kitchen, for I didn't explicitly say the turkey would come along with the plate. One can of course patch up the definition of 'bring' or 'plate' to handle this problem, but only at the cost of creating others. (Will a few more patches tame the problem? At what point should one abandon patches and seek an altogether new approach? Such are the methodological uncertainties regularly encountered in this field, and of course no one can responsibly claim in advance to have a good rule for dealing with them. Premature counsels of despair or calls for revolution are as clearly to be shunned as the dogged pursuit of hopelesss avenues; small wonder the field is contentious.)

While one cannot get away with the tactic of supposing that one can do just one thing, it remains true that very little of what could (logically) happen in any situation does happen. Is there some way of fallibly marking the likely area of important side-effects, and assuming the rest of the situation to stay unchanged? Here is where relevance tests seem like a good idea, and they may well be, but not within the deductive approach. As Minsky notes:

Even if we formulate relevancy restrictions, logistic systems have a problem using them. In any logistic system, all the axioms are necessarily 'permissive'—they all help to permit new inferences to be drawn. Each added axiom means more theorems; none can disappear. There simply is no direct way to add information to tell such a system about kinds of conclusions that should not be drawn! ... If we try to change this by adding axioms about relevancy, we still produce all the unwanted theorems, plus annoying statements about their irrelevancy (Minsky 1981:125).

What is needed is a system that genuinely ignores most of what it knows, and operates with a well-chosen portion of its knowledge at any moment. Well-chosen, but not chosen by exhaustive consideration. How, though, can you give a system rules for ignoring—or better, since explicit rule-following is not the problem, how can you design a system that reliably ignores what it ought to ignore under a wide variety of different circumstances in a complex action environment?

John McCarthy calls this the qualification problem, and vividly illustrates it via the famous puzzle of the missionaries and the cannibals.

Three missionaries and three cannibals come to a river. A rowboat that seats two is available. If the cannibals ever outnumber the missionaries on either bank of the river, the missionaries will be eaten. How shall they cross the river?

Obviously the puzzler is expected to devise a strategy of rowing the boat back and forth that gets them all across and avoids disaster...

Imagine giving someone the problem, and after he puzzles for a while, he suggests going upstream half a mile and crossing on a bridge. 'What bridge?' you say. 'No bridge is mentioned in the statement of the problem.' And this dunce replies, 'Well, they don't say there isn't a bridge.' You look at the English and even at the translation of the English into first order logic, and you must admit that 'they don't say' there is no bridge. So you modify the problem to exclude bridges and pose it again, and the dunce proposes a helicopter, and after you exclude that, he proposes a winged horse or that the others hang onto the outside of the boat while two row.

You now see that while a dunce, he is an inventive dunce. Despairing of getting him to accept the problem in the proper puzzler s spirit, you tell him the solution. To your further annoyance, he attacks your solution on the grounds that the boat might have a leak or lack oars. After you rectify that omission from the statement of the problem. he suggests that a sea monster may swim up the river and may swallow the boat. Again you are frustrated, and you look for a mode of reasoning that will settle his hash once and for all (McCarthy 1980: 29-30).

What a normal, intelligent human being does in such a situation is to engage in some form of non-monotonic inference. In a classical, monotonic logical system, adding premisses never diminishes what can be proved from the premisses. At Minsky noted, the axioms are essentially permissive, and once a theorem is permitted, adding more axioms will never invalidate the proofs of earlier theorems. But when we think about a puzzle or a real-life problem, we can achieve a solution (and even prove that it is a solution, or even the only solution to that problem), and then discover our solution invalidated by the addition of a new element to the posing of the problem; e.g. 'I forgot to tell you—there are no oars' or 'By the way, there's a perfectly good bridge upstream.'

What such late additions show us is that, contrary to our assumption, other things weren't equal. We had been reasoning with the aid of a ceteris paribus assumption, and now our reasoning has just been jeopardized by the discovery that something 'abnormal' is the case. (Note, by the way, that the abnormality in question is a much subtler notion than anything anyone has yet squeezed out of probability theory. As McCarthy notes, 'The whole situation involving cannibals with the postulated properties cannot be regarded as having a probability, so it is hard to take seriously the conditional probability of a bridge given the hypothesis' (ibid.).)

The beauty of a ceteris paribus clause in a bit of reasoning is that one does not have to say exactly what it means. 'What do you mean, "other things being equal"? Exactly which arrangements of which other things count as being equal?' If one had to answer such a question, invoking the ceteris paribus clause would be pointless, for it is precisely in order to evade that task that one uses it. If one could answer that question, one wouldn't need to invoke the clause in the first place. One way of viewing the frame problem, then, is as the attempt to get a computer to avail itself of this distinctively human style of mental operation. There are several quite different approaches to non-monotonic inference being pursued in AI today. They have in common only the goal of capturing the human talent for ignoring what should be ignored, while staying alert to relevant recalcitrance when it occurs.

One family of approaches, typified by the work of Marvin Minsky and Roger Schank (Minsky 1981; Schank and Abelson 1977), gets its ignoring power from the attention-focusing power of stereotypes. The inspiring insight here is the idea that all of life's experiences, for all their variety, boil down to variations on a manageable number of stereotypic themes, paradigmatic scenarios—'frames' in Minsky's terms, 'scripts' in Schank's.

An artificial agent with a well-stocked compendium of frames or scripts, appropriately linked to each other and to the impingements of the world via its perceptual organs, would face the world with an elaborate system of what might be called habits of attention and benign tendencies to leap to particular sorts of conclusions in particular sorts of circumstances. It would, 'automatically' pay attention to certain features in certain environments and assume that certain unexamined normal features of those environments were present. Concomitantly, it would be differentially alert to relevant divergences from the stereotypes it would always begin by 'expecting'.

Simulations of fragments of such an agent's encounters with its world reveal that in many situations it behaves quite felicitously and apparently naturally, and it is hard to say, of course, what the limits of this

approach are. But there are strong grounds for skepticism. Most obviously, while such systems perform creditably when the world co-operates with their stereotypes, and even with anticipated variations on them, when their worlds turn perverse, such systems typically cannot recover gracefully from the misanalyses they are led into. In fact, their behaviour in extremis looks for all the world like the preposterously counter-productive activities of insects betrayed by their rigid tropisms and other genetically hard-wired behavioural routines.

When these embarrassing misadventures occur, the system designer can improve the design by adding provisions to deal with the particular cases. It is important to note that in these cases, the system does not redesign itself (or learn) but rather must wait for an external designer to select an improved design. This process of redesign recapitulates the process of natural selection in some regards; it favours minimal, piecemeal, ad hoc redesign which is tantamount to a wager on the likelihood of patterns in future events. So in some regards it is faithful to biological themes. Nevertheless, until such a system is given a considerable capacity to learn from its errors without designer intervention, it will continue to respond in insect-like ways, and such behaviour is profoundly unrealistic as a model of human reactivity to daily life. The short cuts and cheap methods provided by a reliance on stereotypes are evident enough in human ways of thought, but it is also evident that we have a deeper understanding to fall back on when our short cuts don't avail, and building some measure of this deeper understanding into a system appears to be a necessary condition of getting it to learn swiftly and gracefully.

In effect, the script or frame approach is an attempt to pre-solve the frame problems the particular agent is likely to encounter. While insects do seem saddled with such control systems, people, even when they do appear to be relying on stereotypes, have back-up systems of thought that can deal more powerfully with problems that arise. Moreover, when people do avail themselves of stereotypes, they are at least relying on stereotypes of their own devising, and to date no one has been able to present any workable ideas about how a person's frame-making or script-writing machinery might be guided by its previous experience.

Several different sophisticated attempts to provide the representational framework for this deeper understanding have emerged from the deductive tradition in recent years. Drew McDermott and Jon Doyle have developed a 'non-monotonic logic' (1980), Ray Reiter has a 'logic for default reasoning' (1980), and John McCarthy has developed a system of 'circumscription', a formalized 'rule of conjecture that can be used by a person or program for "jumping to conclusions"' (1980). None of these is, or is claimed to be, a complete solution to the problem of ceteris paribus reasoning, but they might be components of such a solution. More recently, McDermott has offered a 'temporal logic for reasoning about processes and plans' (McDermott 1982). I will not attempt to assay the formal strengths and weaknesses of these approaches. Instead I will concentrate on another worry. From one point of view, non-monotonic or default logic, circumscription, and temporal logic all appear to be radical improvements to the mindless and clanking deductive approach, but from a slightly different perspective they appear to be more of the same, and at least as unrealistic as frameworks for psychological models.

They appear in the former guise to be a step towards greater psychological realism, for they take seriously, and attempt to represent, the phenomenologically salient phenomenon of common sense ceteris paribus 'jumping to conclusions' reasoning. But do they really succeed in offering any plausible suggestions about how the backstage implementation of that conscious thinking is accomplished in people? Even if on some glorious future day a robot with debugged circumscription methods manoeuvred well in a non-toy environment, would there be much likelihood that its constituent processes, described at levels below the phenomenological, would bear informative relations to the unknown lower-level backstage processes in human beings? To bring out better what my worry is, I want to introduce the concept of a cognitive wheel.

We can understand what a cognitive wheel might be by reminding ourselves first about ordinary wheels. Wheels are wonderful, elegant triumphs of technology. The traditional veneration of the mythic inventor of the wheel is entirely justified. But if wheels are so wonderful, why are there no animals with wheels? Why are no wheels to be found (functioning as wheels) in nature? First, the presumption of that question must be qualified. A few years ago the astonishing discovery was made of several microscopic beasties (some bacteria and some unicellular eukaryotes) that have wheels of sorts. Their propulsive tails, long thought to be flexible flagella, turn out to be more or less rigid corkscrews, which rotate continuously', propelled by microscopic motors of sorts, complete with main bearings. Better known, if less interesting for obvious reasons, are the tumbleweeds. So it is not quite true that there are no wheels (or wheeliform designs) in

nature.

Still, macroscopic wheels—reptilian or mammalian or avian wheels - are not to be found. Why not? They would seem to be wonderful retractable landing gear for some birds, for instance. Once the question is posed, plausible reasons rush in to explain their absence. Most important, probably, are the considerations about the topological properties of the axle/bearing boundary that make the transmission of material or energy across it particularly difficult. How could the life-support traffic arteries of a living system maintain integrity across this boundary? But once that problem is posed, solutions suggest themselves; suppose the living wheel grows to mature form in a non-rotating, non-functional form, and is then hardened and sloughed off, like antlers or an outgrown shell, but not completely off: it then rotates freely on a lubricated fixed axle. Possible? It's hard to say. Useful? Also hard to say, especially since such a wheel would have to be free-wheeling. This is an interesting speculative exercise, but certainly not one that should inspire us to draw categorical, a priori conclusions. It would be foolhardy to declare wheels biologically impossible, but at the same time we can appreciate that they are at least very distant and unlikely solutions to natural problems of design.

Now a cognitive wheel is simply any design proposal in cognitive theory (at any level from the purest semantic level to the most concrete level of 'wiring diagrams' of the neurones) that is profoundly unbiological, however wizardly and elegant it is as a bit of technology.

Clearly this is a vaguely defined concept, useful only as a rhetorical abbreviation, as a gesture in the direction of real difficulties to be spelled out carefully. 'Beware of postulating cognitive wheels' masquerades as good advice to the cognitive scientist, while courting vacuity as a maxim to follow. It occupies the same rhetorical position as the stockbroker's maxim: buy low and sell high. Still, the term is a good theme-fixer for discussion.

Many critics of AI have the conviction that any AI system is and must be nothing but a gearbox of cognitive wheels. This could of course turn out to be true, but the usual reason for believing it is based on a misunderstanding of the methodological assumptions of the field. When an AI model of some cognitive phenomenon is proposed, the model is describable at many different levels, from the most global, phenomenological level at which the behaviour is described (with some presumptuousness) in ordinary mentalistic terms, down through various levels of implementation all the way to the level of program code—and even further down, to the level of fundamental hardware operations if anyone cares. No one supposes that the model maps onto the process of psychology and biology all the way down. The claim is only that for some high level or levels of description below the phenomenological level (which merely sets the problem) there is a mapping of model features onto what is being modelled: the cognitive processes in living creatures, human or otherwise. It is understood that all the implementation details below the level of intended modelling will consist, no doubt, of cognitive wheels—bits of unbiological computer activity mimicking the gross effects of cognitive subcomponents by using methods utterly unlike the methods still to be discovered in the brain. Someone who failed to appreciate that a model composed microscopically of cognitive wheels could still achieve a fruitful isomorphism with biological or psychological processes at a higher level of aggregation would suppose there were good a priori reasons for generalized skepticism about AI.

But allowing for the possibility of valuable intermediate levels of modelling is not ensuring their existence. In a particular instance a model might descend directly from a phenomenologically recognizable level of psychological description to a cognitive wheels implementation without shedding any light at all on how we human beings manage to enjoy that phenomenology. I suspect that all current proposals in the field for dealing with the frame problem have that shortcoming. Perhaps one should dismiss the previous sentence as mere autobiography. I find it hard to imagine (for what that is worth) that any of the procedural details of the mechanization of McCarthy's circumscriptions, for instance, would have suitable counter-parts in the backstage story yet to be told about how human common-sense reasoning is accomplished. If these procedural details lack 'psychological reality' then there is nothing left in the proposal that might model psychological processes except the phenomenological-level description in terms of jumping to conclusions, ignoring, and the like—and we already know we do that.

There is an alternative defence of such theoretical explorations, however, and I think it is to be taken seriously. One can claim (and I take McCarthy to claim) that while formalizing common-sense reasoning in his fashion would not tell us anything directly about psychological processes of reasoning, it would clarify, sharpen, systematize the purely semantic-level characterization of the demands on any such implementation,

biological or not. Once one has taken the giant step forward of taking information-processing seriously as a real process in space and time, one can then take a small step back and explore the implications of that advance at a very abstract level. Even at this very formal level, the power of circumscription and the other versions of non-monotonic reasoning remains an open but eminently explorable question.

Some have thought that the key to a more realistic solution to the frame problem (and indeed, in all likelihood, to any solution at all) must require a complete rethinking of the semantic-level setting, prior to concern with syntactic-level implementation. The more or less standard array of predicates and relations chosen to fill out the predicate-calculus format when representing the 'propositions believed' may embody a fundamentally inappropriate parsing of nature for this task. Typically, the interpretation of the formulae in these systems breaks the world down along the familiar lines of objects with properties at times and places. Knowledge of situations and events in the world is represented by what might be called sequences of verbal snapshots. State S, constitutively described by a list of sentences true at time t asserting various n-adic predicates true of various particulars, gives way to state S', a similar list of sentences true at t'. Would it perhaps be better to reconceive of the world of planning in terms of histories and processes? Instead of trying to model the capacity to keep track of things in terms of principles for passing through temporal cross-sections of knowledge expressed in terms of terms (names for things, in essence) and predicates, perhaps we could model keeping track of things more directly, and let all the cross-sectional information about what is deemed true moment by moment be merely implicit (and hard to extract—as it is for us) from the format. These are tempting suggestions, but so far as I know they are still in the realm of handwaving.

Another, perhaps related, handwaving theme is that the current difficulties with the frame problem stem from the conceptual scheme engendered by the serial-processing von Neumann architecture of the computers used to date in AI. As large, fast parallel processors are developed, they will bring in their wake huge conceptual innovations which are now of course only dimly imaginable. Since brains are surely massive parallel processors, it is tempting to suppose that the concepts engendered by such new hardware will be more readily adaptable for realistic psychological modelling. But who can say? For the time being, most of the optimistic claims about the powers of the parallel-processing belong in the same camp with the facile observations often encountered in the work of neuroscientists, who postulate marvellous cognitive powers for various portions of the nervous system without a clue how they are realized.

Filling in the details of the gap between the phenomenological magic show and the well-understood powers of small tracts of brain tissue is the immense research task that lies in the future for theorists of every persuasion. But before the problems can be solved they must be encountered, and to encounter the problems one must step resolutely into the gap and ask how-questions. What philosophers (and everyone else) have always known is that people—and no doubt all intelligent agents—can engage in swift, sensitive, risky-but-valuable ceteris paribus reasoning. How do we do it? AI may not yet have a good answer, but at least it has encountered the question.

REFERENCES

Cherniak, C. [1983]. 'Rationality and the Structure of Memory.' Synthese (57:163-86).

Darmstadter, H. (1971). 'Consistency of Belief.' J. Philosophy 68: 301-10.

Dennett, D. C. (1978). Brainstorms . Cambridge, Mass.: MIT Press/Bradford Books.

Dennett, D. C. (1982a). 'Why Do We Think What We Do About Why We Think What We Do?', Cognition 12: 219-27.

Dennett, D. C. (19826). 'How to Study Consciousness Empirically; Or Nothing Comes to Mind.', Synthese 53:159-80.

Dennett, D. C. (1982c). 'Beyond Belief.' In A. Woodfield (ed.). Thought and Object, pp.1-96. Oxford: Clarendon Press.

Dennett, D. C. (1983). 'Styles of Mental Representation.' Proc. Aristotelian Soc. 83: 213-26.

Diamond, J. (1983). 'The Biology of the Wheel.' Nature 302: 572-3.

Dreyfus, H. L. (1972). What Computers Can't Do. New York: Harper & Row.

Fikes, R., and Nilsson, N. (1971). 'STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving,' Artificial Intelligence 2:189-208.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston, Mass.: Houghton-Mifflin.

Goodman, N. (1965). Fact, Fiction and Forecast, 2nd edn. Indianapolis: Bobbs?Merrill.

Goodman, N. (1982). 'Thoughts Without Words.', Cognition 12: 211-17.

Hayes, P.J. (1978). 'Naive Physics I: The Ontology of Liquids.' Working Paper 35, Institute for Semantic and Cognitive Studies, Geneva.

Hayes, P.J. (1979). 'The Naive Physics Manifesto.' In D. Michie (ed.) Expert Systems in the Micro-Electronic Age, pp. 242-70. Edinburgh: Edinburgh University Press.

Hofstadter, D. (1982). 'Can Inspiration be Mechanized?' Scientific American 247: 18-34.

McCarthy, J. (1968). 'Programs with Common Sense.' Proceedings of the Teddington Conference on the Mechanization of Thought Processes, London. Repr. in M. Minsky (ed.)., Semantic Information Processing, pp. 40 -58. Cambridge, Mass.: MIT Press.

McCarthy, J. (1980). 'Circumscription—A Form of Non-Monotonic Reasoning.' Artificial Intelligence 13: 27-39.

McCarthy, J. and Hayes, P. J. (1969). 'Some Philosophical Problems from the Standpoint of Artificial Intelligence.' In B. Meltzer and D. Michie (eds.), Machine Intelligence 4, pp. 463-502. Edinburgh: Edinburgh University Press.

McDermott, D. (1982). 'A Temporal Logic for Reasoning about Processes and Plans.' Cognitive Science 6:101-55.

McDermott, D. and Doyle, J. (1980). 'Non-Monotonic Logic.' Artifical Intelligence 13: 41- 72.

Millikan, R. G. [1984]. Language, Thought and Other Biological Categories. Cambridge, Mass.: MIT Press/Bradford Books.

Minsky, M. (1981). 'A Framework for Representing Knowledge.' Originally published as MIT AI Lab. Memo 3306. Quotation drawn from excerpts repr. in J. Haugeland (ed.), Mind Design, pp. 95-128. Cambridge, Mass.: MIT Press, Bradford Books.

Newell, A. (1982). 'The Knowledge Level.' Artificial Intelligence 15: 87-127.

Reiter, R. (1980). 'A Logic for Default Reasoning.' Artificial Intelligence 13: 81-132.

Ryle, G. (1949). The Concept of Mind. London: Hutchinson.

Schank, R. C., and Abelson, R. P. (1977). Scripts, Plans, Goals, and Understanding : An Inquiry into Human Knowledge. Hillsdale NJ: Erlbaum.

Selfridge, O. (forthcoming). Tracking and Trailing. Cambridge, Mass.: MIT Press, Bradford Books.

de Sousa, R. (1979). 'The Rationality of Emotions. J. Dialogue 18: 41-63.

Wooldridge, D. (1963). The Machinery of the Brain. New York: MeGraw-Hill.