Artificial Intelligence 47 (1991) 3–30.
Elsevier

Foundations of AI: the big issues*

David Kirsh

Department of Cognitive Science C-015, University of California, San Diego, La Jolla, CA 92093, USA

Received December 1989
Revised October 1990

Abstract

Kirsh, D., Foundations of AI: the big issues, Artificial Intelligence 47 (1991) 3–30.

The objective of research in the foundations of AI is to explore such basic questions as: What is a theory in AI? What are the most abstract assumptions underlying the competing visions of intelligence? What are the basic arguments for and against each assumption? In this essay I discuss five foundational issues: (1) Core AI is the study of conceptualization and should begin with knowledge level theories. (2) Cognition can be studied as a disembodied process without solving the symbol grounding problem. (3) Cognition is nicely described in propositional terms. (4) We can study cognition separately from learning. (5) There is a single architecture underlying virtually all cognition. I explain what each of these implies and present arguments from both outside and inside AI why each has been seen as right or wrong.

1. Introduction

In AI, to date, there has been little discussion, and even less agreement, on methodology: What is a theory in AI? An architecture? An account of knowledge? Can a theory be tested by studying performance in abstract, simulated environments, or is it necessary to hook up implementations to actual visual input and actual motor output? Is there one level of analysis or a small set of problems which ought to be pursued first? For instance, should we try to identify the knowledge necessary for a skill before we concern ourselves with issues of representation and control? Is complexity theory relevant to the central problems of the field? Indeed, what are the central problems?

The objective of research in the foundations of AI is to address some of

* Support for this work has been provided in part by the Army Institute for Research in Management, Information and Communication Systems contract number DAKF11-88-C-0045.

these basic questions of method, theory and orientation. It is to self-consciously reappraise what AI is all about.

The pursuit of AI does not occur in isolation. Fields such as philosophy, linguistics, psychophysics and theoretical computer science have exercised a historical influence over the field and today there is as much dialogue as ever, particularly with the new field of cognitive science. One consequence of dialogue is that criticisms of positions held in one discipline frequently apply to positions held in other disciplines.

In this first essay, my objective is to bring together a variety of these arguments both for and against the dominant research programs of AI.

It is impossible, of course, to explore carefully all of these arguments in a single paper. The majority, in any event, are discussed in the papers in this volume, and it is not my intent to repeat them here. It may be of use, though, to stand back and consider several of the most abstract assumptions underlying the competing visions of intelligence. These assumptions—whether explicitly named by theorists or not—identify issues which have become focal points of debate and serve as dividing lines of positions.

Of these, five stand out as particularly fundamental:

- *Pre-eminence of knowledge and conceptualization*: Intelligence that transcends insect-level intelligence requires declarative knowledge and some form of reasoning-like computation—call this *cognition*.[1] Core AI is the study of the conceptualizations of the world presupposed and used by intelligent systems during cognition.
- *Disembodiment*: Cognition and the knowledge it presupposes can be studied largely in abstraction from the details of perception and motor control.
- *Kinematics of cognition are language-like*: It is possible to describe the trajectory of knowledge states or informational states created during cognition using a vocabulary very much like English or some regimented logico-mathematical version of English.
- *Learning can be added later*: The kinematics of cognition and the domain knowledge needed for cognition can be studied separately from the study of concept learning, psychological development, and evolutionary change.
- *Uniform architecture*: There is a single architecture underlying virtually all cognition.

Different research programs are based, more or less, on an admixture of these assumptions plus corollaries.

[1] By cognition I do not mean to take a stand on what the proper subject matter of cognitive science is. The term is meant to refer to computational processes that resemble both reasoning in a classical sense and computational processes that are more "peripheral" than reasoning, such as language recognition and object identification, where the representations are not about the entities and relations we have common sense terms for, but which may still usefully be construed as rules operating on representations.

Logicism [15, 32] as typified by formal theorists of the commonsense world, formal theorists of language and formal theorists of belief [17, 24], presupposes almost all of these assumptions. Logicism, as we know it today, is predicated on the pre-eminence of reasoning-like processes and conceptualization, the legitimacy of disembodied analysis, on interpreting rational kinematics as propositional, and the possibility of separating thought and learning. It remains neutral on the uniformity of the underlying architecture.

Other research progams make a virtue of denying one or more of these assumptions. Soar, [30, 35] for instance, differs from logicism in according learning a vital role in the basic theory and in assuming that all of cognition can be explained as processes occurring in a single uniform architecture. Rational kinematics in Soar are virtually propositional but differ slightly in containing control markers—preferences—to bias transitions. In other respects, Soar shares with logicism the assumption that reasoning-like processes and conceptualization are central, and that it is methodologically acceptable to treat central processes in abstraction from perceptual and motor processes.

Connectionists, [27, 38] by contrast, deny that reasoning-like processes are pre-eminent in cognition, that core AI is the study of the concepts underpinning domain understanding, and that rational kinematics is language-like. Yet like Soar, connectionists emphasize the centrality of learning in the study of cognition, and like logicists they remain agnostic about the uniformity of the underlying architecture. They are divided on the assumption of disembodiment.

Moboticists [3] take the most extreme stance and deny reasoning, conceptualization, rational kinematics, disembodiment, uniformity of architecture and the separability of knowledge and learning (more precisely evolution). Part of what is attractive in the mobotics approach is precisely its radicalness.

Similar profiles can be offered for Lenat and Feigenbaum's position [23], Minsky's society of mind theory [28], Schank's anti-formalist approach [40, 41] and Hewitt and Gasser's account [12, 14] of much of distributed AI research.

These five issues by no means exhaust the foundational issues posed by the various approaches. But each does, in my opinion, lie at the center of a cluster of deep questions.

In what follows I will explore arguments for and against each of these assumptions. I will explain what each of them implies and why they have been seen as right or wrong.

2. Are knowledge and conceptualization at the heart of AI?

Here is one answer to the question: what is a theory in AI?

> A theory in AI is a specification of the knowledge underpinning a cognitive skill.

A cognitive skill is the information-based control mechanism regulating performance in some domain. It is meant to cover the gamut of information-sensitive activities such as problem solving, language use, decision making, routine activity, perception and some elements of motor control.

In accepting the priority of knowledge level theories, one is not committed to supposing that knowledge is explicitly encoded declaratively and deployed in explicitly inferential processes, although frequently knowledge will be. One's commitment is that knowledge and conceptualization lie at the heart of AI: that a major goal of the field is to discover the basic knowledge units of cognition (of intelligent skills).

What are these knowledge units? In the case of qualitative theories of the commonsense world, and in the case of Lenat's CYC project [21, 23], these basic knowledge units are the conceptual units of *consensus reality*—the core concepts underpinning "the millions of things that we all know and that we assume everyone else knows" [21, p. 4]. Not surprisingly, these concepts are often familiar ideas with familiar names—though sometimes they will be theoretical ideas, having a technical meaning internal to the theory. For instance, in CYC, in addition to terms for tables, salt, Africa, and numbers—obvious elements of consensual reality—there are technical terms such as temporal subabstraction, temporal projectability, partition, change predicate which have no simple correlate in English, and which are included as abstract elements of consensual reality because of the difficulty of constructing an adequate account without them.

In the case of linguistics and higher vision these basic knowledge units tend more generally to be about theoretical entities. Only occasionally will there be pre-existing terms in English for them. Thus, noun phrase, sphere, pyramid and other shapes are commonsense concepts having familiar English names, but governing domain, animate movements, causal launchings[2] and most shape representations are, for most people, novel ideas that are not part of common parlance. The basic knowledge units of cognition—the conceptualizations underpinning cognitive skills—may range, then, from the familiar to the exotic and theoretical.

The basic idea that knowledge and conceptualization lie at the heart of AI stems from the seductive view that cognition is inference. Intelligent skills, an old truism of AI runs, are composed of two parts: a declarative knowledge base and an inference engine.

The inference engine is relatively uncomplicated; it is a domain-independent program that takes as input a set of statements about the current situation plus a fragment of the declarative knowledge base, it produces as output a stream of

[2] It is widely argued in the developmental literature that one of the earliest and visually most robust cues for distinguishing animate creatures like dogs and snakes from non-animate objects like toy dogs, and cars, which may also move, are cues about body part trajectories, and original causation [25].

inferred declaratives culminating in the case of decision making and routine activity, in directives for appropriate action.

In contrast to the inference engine, the knowledge base is domain-specific and is as complicated as a cognitive skill requires. Domain knowledge is what distinguishes the ability to troubleshoot a circuit from the ability to understand the meaning of a sentence. Both require knowledge but of different domains. It follows that the heart of the AI problem is to discover what an agent knows about the world which permits success. This idea, *in one form or another*, has been endorsed by logicists, by Lenat and Feigenbaum [23], Chomsky [6], Montague [29], and with variations by Schank [41], and Newell and Simon [32].

The qualification *in one form or another* is significant. As mentioned, a commitment to theorizing about knowledge and knowledge units is not in itself a commitment to large amounts of on-line logical reasoning or explicit representation of domain knowledge. It is well known that not all skills that require intelligent control require an *explicit* knowledge base. So it is a further thesis that declarative knowledge and logical inference are actually deployed in most cognitive skills. In such cases we still may say that cognition is inference, but we no longer expect to find explicit inference rules or even complete trajectories of inferential steps. In the source code of cognition we would find instructions for inferential processes throughout. But knowledge can be compiled into procedures or designed into control systems which have no distinct inference engines. So often our account of cognition is more of the form "The system is acting *as if* it were inferring . . .".

Knowledge compilation One question of considerable interest among theorists who accept the centrality of knowledge and the virtue of knowledge level theories, is "How far can this knowledge compilation go?"

According to Nilsson there are severe limits on this compilation. Overt declaratives have special virtues.

> The most versatile intelligent machines will represent much of their knowledge about their environment declaratively . . . [A declarative can] be used by the machine even for purposes unforeseen by the machine's designer, it [can] more easily be modified than could knowledge embodied in programs, and it facilitate[s] communication between machine and other machines and humans. [33]

For Nilsson, the theory of what is known is a good approximation of what is actually represented declaratively. He suggests that some reactions to situations and some useful inferences may be compiled. But storage and indexing costs militate against compiling knowledge overmuch. Real flexibility requires explicit declarative representation of knowledge. No doubt, it is an empirical question just how much of a cognitive skill can be compiled. But as long as a system uses some explicit declaratives, the apparatus of declarative representation must be in place, making it possible, when time permits, to control action through run time inference.

Rosenschein et al. [37] see the inflexibility of knowledge compilation as far less constraining. On their view, a significant range of tasks connected with adaptive response to the environment can be compiled. To determine the appropriate set of reactions to build into a machine, a designer performs the relevant knowledge level logical reasoning at compile time so that the results will be available at run time. Again, it is an empirical matter how many cognitive skills can be completely automatized in this fashion. But the research program of situated automata is to push the envelope as far as possible.

A similar line of thought applies to the work of Chomsky and Montague. When they claim to be offering a theory about the knowledge deployed in parsing and speech production it does not follow they require on-line inference. By offering their theories in the format of "here's the knowledge base use the obvious inference engine" they establish the effectiveness of their knowledge specification: it is a condition on their theory that when conjoined with the obvious inference engine it should generate all and only syntactic strings (or some specified fragment of that set). That is why their theories are called *generative*. But to date no one has offered a satisfactory account of how the theory is to be efficiently implemented. Parsing *may* involve considerable inference, but equally it may consist of highly automated retrieval processes where structures or fragments of structures previously found acceptable are recognized. To be sure, some theorists say that recognition is itself a type of inference: that recognizing a string of words *as* an NP involves inference. Hence even parsing construed as constraint satisfaction or as schema retrieval (instantiation) and so forth, is itself inferential at bottom. But this is not the dominant view. Whatever the answer, though, there are no *a priori* grounds for assuming that statements of linguistic principle are encoded explicitly in declaratives and operated on by explicit inference rules.

Whether knowledge be explicit or compiled, the view that cognition is inference and that theorizing at the *knowledge level* is at least the starting place of scientific AI is endorsed by a large fragment of the community.

Opposition In stark contrast is the position held by Rod Brooks. According to Brooks [3] a theory in AI is not an account of the knowledge units of cognition. Most tasks that seem to involve considerable world knowledge may yet be achievable without appeal to declaratives, to concepts, or to basic knowledge units, even at compile time. Knowledge level theories, he argues, too often chase fictions. If AI's overarching goal is to understand intelligent control of action, then if it turns out to be true, as Brooks believes it will, that most intelligent behaviour can be produced by a system of carefully tuned control systems interconnected in a simple but often ad hoc manner, then why study knowledge? A methodology more like experimental engineering is what is required.

If Brooks is right, intelligent control systems can be designed before a designer has an articulated conceptualization of the task environment. Moreover, the system itself can succeed without operating on a conceptualization in any interesting sense. New behaviours can be grown onto older behaviours in an evolutionary fashion that makes otiose the task of conceptualizing the world. The result is a system that, for a large class of tasks, might match the versatility of action achievable with declaratives, yet it never calls on the type of capacities we associate with having knowledge of a conceptualization and symbolic representation of basic world elements.

Whatever our belief about the viability of Brooks' position he has succeeded in exposing an important foundational question: *Why assume intelligence requires concepts*? If the AI community has largely ignored this problem it is not simply because it is a presupposition of the view that cognition is inference. It is also because the problem of designing intelligent systems has never been consciously formulated as one of discovering concepts in a *psychological* sense. In AI there is no marked difference between assuming a system to have a symbol in a declarative and assuming it to have a concept. The worry about what it is to have a concept is seldom articulated. Hence skepticism about concepts and conceptualization has been kept down.

2.1. *Are concepts really necessary for most intelligence?*

Evidence that the notion of concept is understudied in AI is easy to find. When Nilsson, for instance, unambiguously states that "The most important part of the 'AI problem' involves inventing an appropriate conceptualization" [33, p. 10], it would be natural to expect him to provide an account of what it is for a system to have a concept. But in fact by conceptualization he does not mean the concepts a system has about the world. Rather he means the *designer* of a machine's best guess about a "mathematical structure consisting of objects, functions, and relations" close enough to the real world for the machine to achieve its purposes. Admittedly, for Nilsson, the designer *builds his conceptualization into* a system by creating "linguistic terms to denote his invented objects, functions and relations", putting these terms in sentences in the predicate calculus, and giving "the machine declarative knowledge about the world by storing these sentences in the machine's memory". So in certain cases talk of conceptualization is short hand for talk of the concepts a machine has. But it is important to mark the logical distinction between:

(1) the conceptualization of a task the designer has;
(2) the conceptual system the machine embodying the skill has;
(3) the way the conceptual system is encoded.

The difference lies in the deeply philosophical question of what it is to *grasp* a concept. We cannot just assume that a machine which has a structure in memory that corresponds in name to a structure in the designer's conceptualization is sufficient for grasping the concept. The structure must play a role in a network of abilities; it must confer on the agent certain causal powers [1]. Some of these powers involve reasoning: being able to use the structure *appropriately* in deduction, induction and perhaps abduction. But other powers involve perception and action—hooking up the structure via causal mechanisms to the outside world.

Logicists are not unmindful of the need to explain what it is for a system to understand a proposition, or to grasp the concepts which constitute propositions. But the party line is that this job can be pursued independently from the designer's main task of inventing conceptualizations. The two activities—inventing conceptualizations and grounding concepts—are modular. Hence the grounding issue has not historically been treated as posing a challenge that might overturn the logicist program.

A similar belief in modularizing the theorist's job is shared by Lenat and Feigenbaum. They see the paramount task of AI to be to discover the conceptual knowledge underpinning cognitive skills and consensus reality. This leaves open the question of what exactly grasping a basic conceptual or knowledge unit of consensus reality amounts to. There certainly is a story of grounding to be told, but creatures with different perceptual-motor endowments will each require its own story. So why not regard the problem of conceptualization to be independent from the problem of grounding concepts?

This assumption of modularization—of disembodiment—is the core concern of Brian Smith [42] in his reply to Lenat and Feigenbaum. It pertains, as well, to worries Birnbaum expresses about model theoretic semantics [1]. Both Birnbaum and Smith emphasize that if knowing a concept, or if having knowledge about a particular conceptualization requires a machine to have a large background of behavioural, perceptual and even reasoning skills, then the greater part of the AI task may reside in understanding how concepts can refer, or how they can be used in reasoning, perceiving, acting, rather than in just identifying those concepts or stating their axiomatic relations.

Accordingly, it is time to explore what the logicist's conception of a concept amounts to. Only then can we intelligently consider whether it is fair to say that logicists and Lenat and Feigenbaum—by assuming they can provide a machine with symbols that are not *grounded* and so not truly grasped—are omitting an absolutely major part of the AI problem.

2.1.1. The logicist concept of concept

A concept, on anyone's view, is a modular component of knowledge. If we say John knows *the pen is on the desk*, and we mean this to imply that John grasps the fact of there being a particular pen on a particular desk, we assume that he has distinct concepts for *pen*, *desk* and *on*. We assume this because we believe that John must know what it is for something to be a pen, a desk, and something to be on something else. That is, we assume he has the referential apparatus to think about pens, desks, and being on. At a minimum, this implies having the capacity to substitute other appropriate concepts for x and y in (*On pen y*), (*On x desk*), and R in (*R pen desk*). If John could not just as easily understand what it is for a pen to be on something other than a desk, or a desk to have something other than a pen on it, he would not have enough understanding of *pen*, *desk*, and *on* to be able to display the minimal knowledge that pens and desks are distinct entities with enough causal individuality to appear separately, and in different combinations.

Now the basic premiss driving the logicist program, as well as Lenat and Feigenbaum's search for the underpinnings of consensus reality, is that to understand an agent's knowledge we must discover the structured system of concepts underpinning its skills. This structure can be discovered without explaining all that is involved in having the *referential apparatus* presupposed by concepts because it shows up in a number of purely disembodied, rational processes. If concepts and conceptual schemes seem to play enough of an explanatory role at the disembodied level to be seen as robust entities, then we can study their structure without concern for their grounding.

What then are these disembodied processes which can be explained so nicely by disembodied concepts? In the end we may decide that these do not sufficiently ground concepts. But it is important to note their variety. For too often arguments about grounding do not adequately attend to the range of phenomena explained by assuming modular concepts.

Inferential abilities First, and most obviously, is the capacity of an agent to draw inferences. For instance, given the premises that the pen is on the desk, that the pen is matte black, then a knowledgeable agent ought to be able to infer that the matte black pen is on the desk. It often happens that actual agents will not bother to draw this inference. But it is hard for us to imagine that they might have a grasp of what pens are etc, and not be *able* to draw it. Inferences are permissive not obligatory. Thus, as long as it makes sense to view agents to be *sometimes* drawing inferences about a domain, or performing reason-like operations, it makes sense to suppose they have a network of concepts which structures their knowledge.[3]

[3] The much discussed attribute of systematicity which Fodor and Pylyshyn cite in [11] as essential to symbolic reasoning and antithetical to the spirit of much connectionist work to date, is a version of this *generality constraint* on concepts. A few years earlier, Gareth Evans put the matter like this:

> If the subject can be credited with the thought that *a* is *F*, then he must have conceptual resources for entertaining the thought that *a* is *G*, for every property of being *G* of which he has a conception. We thus see the thought that *a* is *F* as lying at the intersection of two series of thoughts: on the one hand, the series of thoughts that *a* is *F*, *b* is *F*, *c* is *F*, . . ., and, on the other hand, the series of thoughts that *a* is *F*, *a* is *G*, *a* is *H*, ... [8, p. 104, footnote 22].

It must be appreciated, however, that when we say that John has the concepts of pen and desk we do not mean that John is able to draw inferences about pens and desks in only a few contexts. He must display his grasp of the terms extensively, otherwise we cannot be sure that he means *desk* by "desk" rather than *wooden object*, for instance. For this reason, if we attribute to a machine a grasp of a single concept we are obliged to attribute it a grasp of a whole system of concepts to structure its understanding. Otherwise its inferential abilities would be too spotty, displaying too many gaps to justify our attribution of genuine understanding. Experience shows that to prevent ridiculous displays of irrationality it is necessary to postulate an elaborate tissue of underlying conceptualizations and factual knowledge. The broader this knowledge base the more robust the understanding, and more reasonable the action. This is one very compelling reason for supposing that intelligence can be studied from a disembodied perspective.

Inferential breadth is only one of the rational capacities that is explained by assuming intelligent agents have concepts. Further capacities include identification and visual attention, learning, knowledge decay and portability of knowledge.

Knowledge and perception Kant once said, sensation without conception is blind. What he meant is that I do not know *what* I am seeing, if I have no concept to categorize my experience. Much of our experience is of a world populated with particular objects, events and processes. Our idea of these things may be abstractions—constructions from something more primitive, or fictional systematizers of experience. But if so, they are certainly robust abstractions, for they let us predict, retrodict, explain and plan events in the world.

It is hard to imagine how we could identify entities if we did not have concepts. The reason this is hard, I suspect, is because object identification is such an active process. Perception, it is now widely accepted, is not a passive system. It is a method for *systematically* gathering evidence about the environment. We can think of it as an oracle offering answers to questions about the external world. Not direct answers, but partial answers, perceptual answers, that serve as evidence for or against certain perceptual *conjectures*. One job of the perceptual system is to ask the right questions. Our eyes jump about an image looking for clues of identity; then shortly thereafter they search for confirmation of conjectures. The same holds for different modalities. Our eyes often confirm or disconfirm what our ears first detect. The notions of evidence, confirmation and falsification, however, are defined as relations between statements or propositions. Concepts are essential to perception then because perception provides evidence for conjectures about the world. It follows that the output of perception must be sufficiently evidence-like—that is, propositional—to be assigned a conceptual structure. How else could we see physical facts, such as the pen being on the desk *as* the structured facts—$|\textit{the pen}|^\frown|\textit{is on}|^\frown|\textit{the desk}|$?

Growth of knowledge A third feature of rational intelligence—learning—can also be partly explained if we attribute to a system a set of disembodied concepts. From the logicist perspective, domain knowledge is much like a theory, it is a system of axioms relating basic concepts. Some axioms are empirical, others are definitional. Learning, on this account, is construed as movement along a trajectory of theories. It is conceptual advance. This approach brings us no closer to understanding the principles of learning, but we have at least defined what these principles are: principles of conceptual advance. A theory of intelligence which did not mention concepts would have to explain learning as a change in capacities behaviourally or functionally classified. Since two creatures with slightly different physical attributes would not have identical capacities, behaviourally defined, the two could not be said to learn identically. Yet from a more abstract perspective, what we are interested in is their knowledge of the domain, the two might indeed seem to learn the same way. Without concepts and conceptual knowledge it is not clear this similarity could be discovered, let alone be explained. But again the relevant notion of concept is not one that requires our knowing how it is grounded. Disembodied concepts serve well enough.

Decay of knowledge In a similar fashion, if a system has a network of disembodied concepts we can often notice and then later explain regularities in how its rational performance degrades. It is an empirical fact that knowledge and skill sometimes decay in existing reasoning systems, such as humans or animals, in a regular manner. Often it does not. Alzheimer's disease may bring about a loss of functionality that is sporadic or at times random. But often, when a system decays, deficits which at first seem to be unsystematic, can eventually be seen to follow a pattern, once we know the structure of the larger system from which they emerge. This is obviously desirable if we are cognitive scientists and wish to explain deficits and predict their etiology; but it is equally desirable if we are designers trying to determine why a design is faulty. If we interpret a system as having a network of concepts we are in a better position to locate where its bugs are. But the fact that we *can* track and *can* explain decay at the conceptual level without explaining grounding offers us further evidence of the robustness of disembodied concepts.

Portability of knowledge There is yet a fifth phenomenon of rationality which the postulation of disembodied concepts can help explain. If knowledge consists in compositions of concepts—that is, propositions—we have an explanation of why, in principle, any piece of knowledge in one microtheory can be combined with knowledge drawn from another microtheory. They can combine because they are structured in a similar fashion out of similar types of elements. At the object level, this explains how it is possible for a cognizer to receive generally useful information in one context, say astronomy, and end up using it in another, say calendar making. At the metalevel, it explains how, as designers, we can build on knowledge in different domains, thereby simplifying our overall account of the knowledge a system requires. Many of the decisions we make rely on information drawn from disparate domains. Knowledge which accrues in one domain can be useful in making decisions in another. This is a fact which Nilsson rightly emphasizes in his condition on portability as a hallmark of commonsense knowledge. Compositionality would explain portability.[4]

Given the virtues of concepts it is hard to imagine anyone seriously doubting that concepts—whose grounding we have yet to explain—lie at the heart of intelligence. Explanations of a system's conceptual system are clearly not the whole story of AI, but can it be reasonably denied that they are a cleanly modular major chapter?

I now turn to these reasonable doubts.

3. Are cognitive skills disembodied?

I have been presenting a justification for the view that, in the main, intelligence can be fruitfully studied on the assumption that the problems and tasks facing intelligent agents can be formally specified, and so pursued abstractly at the knowledge or conceptual level. For analytic purposes we can ask questions about cognitive skills using symbolic characterizations of the environment as input and symbolic characterizations of motor activity as output. Concerns about how conceptual knowledge is *grounded* in perceptual-motor skills can be addressed separately. These questions can be bracketed because what differentiates cognitive skills is not so much the perceptual-motor parameters of a task but the knowledge of the task domain which directs action in that domain. This is the methodological assumption of disembodiment. What are the arguments against it?

In his attack on core AI, Brooks identifies three assumptions related to disembodiment which, in his opinion, dangerously bias the way cognitive skills are studied:

[4] To be sure, this common language of concepts does not apply to *every* domain of knowledge. Microtheories about syntax and early vision, arguably are about domain elements not found in other microtheories. To the degree that the conceptual elements we attribute to syntax and early vision are inaccessible to other inferential processes we are justified in being skeptical of their robustness as concepts in the full blooded sense we mean when we talk of publicly shared concepts like chairs and tables. This concern that we should reserve the term concept for post-peripheral processes is discussed by Cussins [7].

- The output of vision is conceptualized and so the interface between perception and "central cognition" is clean and neatly characterizable in the language of predicate calculus, or some other language with terms denoting objects and terms denoting properties.
- Whenever we exercise our intelligence we call on a central representation of the world state where some substantial fraction of the world state is represented and regularly updated perceptually or by inference.
- When we seem to be pursuing our tasks in an organzied fashion our actions have been planned in advance by envisioning outcomes and choosing a sequence that best achieves the agent's goals.

The error in each of these assumptions, Brooks contends, is to suppose that the real world is somehow simple enough, sufficiently decomposable into concept-sized bites, that we can represent it, in real time, in all the detailed respects that might matter to achieving our goals. It is not. Even if we had enough concepts to cover its relevant aspects we would never be able to compute an updated world model in real time. Moreover, we don't need to. Real success in a causally dense world is achieved by tuning the perceptual system to *action-relevant* changes.

To take an example from J.J. Gibson, an earlier theorist who held similar views, if a creature's goals are to avoid obstacles on its path to a target, it is not necessary for it to constantly judge its distance from obstacles, update a world model with itself at the origin, and recalculate a trajectory given velocity projections. It can instead exploit the invariant relation between its current velocity and instantaneous time to contact obstacles in order to determine a new trajectory directly. It adapts its actions to changes in time to contact. If the environment is perceived in terms of actions that are *afforded* rather than in terms of objects and relations, the otherwise computationally intensive task is drastically simplified.

Now this is nothing short of a Ptolemaic revolution. If the world is always sensed from a perspective which views the environment as *a space of possibilities for action*, then every time an agent performs an action which changes the action potentials which the world affords it, it changes the world as it perceives it. In the last example, this occurs because as the agent changes its instantaneous speed and direction it may perceive significant changes in environmental affordances despite being in almost the same spatial relations to objects in the environment. Even slight actions can change the way a creature perceives the world. If these changes in perception regularly simplify the problem of attaining goals, then traditional accounts of the environment as a static structure composed of objects, relations and functions, may completely misstate the actual computational problems faced by creatures acting in the world. The real problem must be defined relative to the world-for-the-agent. The world-for-the-agent changes despite the world-in-itself remaining constant.

To take another example of how action and perception are intertwined, and so must be considered when stating the computational problems facing agents, consider the problem of grasp planning. Traditionally the problem is defined as follows: Given a target object, an initial configuration of hand joints and free space between hand and target, find a trajectory of joint changes that results in a stable grasp. At one time it was thought that to solve this problem it was necessary to compute the 3D shape of the target, the final configuration of joints, and the trajectory of joint changes between initial and final configurations—a substantial amount of computation by anyone's measure. Yet this is not the problem if we allow compliance. Instead we simply need locate a rough center of mass of the target, send the palm of the hand to that point with the instruction to close on contact, and rely on the hand to *comply* with the object. The problem is elegantly simplified. No longer must we know the shape of the object, the mapping relation between 3D shape and joint configuration, or the constraints on joint closure. The original definition of the grasp planning problem was a mis-statement. It led us to believe that certain subproblems and certain elements of knowledge would be required, when in fact they are not. Compliance changes everything. It alters the way the world should be interpreted.

The point is that the possibility of complying with shapes restructures the world. A creature with a compliant hand confronts a different world than a creature without. Accordingly, a knowledge level account of grasping which did not accommodate the simplifications due to compliance would be false. It would be working with an incorrect set of assumptions about the manipulator.

By analogy, one cardinal idea of the embodied approach to cognition, is that the hardware of the body—in particular, the details of the sensori-motor system—when taken in conjuction with an environment and goals shape the kinds of problems facing an agent. These problems in turn shape the cognitive skills agents have. Consequently, to specify these skills correctly it is necessary to pay close attention to the agent's interactions with its environment—to the actions it does and can do at any point. Disembodied approaches do not interpret the environment of action in this dynamic manner, and so inevitably give rise to false problems and false solutions. They tend to define problems in terms of task environments specified in the abstract perspective independent language of objects and relations.[5]

Now this argument, it seems to me, is sound. But how far does it go? It serves as a reminder to knowledge level theorists that they may easily misspecify a cognitive skill, and that to reliably theorize at the knowledge level one should often have a model of the agent's sensori-motor capacities. But it is

[5] Newell and Simon in their characterization of task environment emphasize that a given physical environment becomes a task environment only relative to a goal or task, and a set of actions. But one assumption they retain is that actions are basically STRIPS-like: they add or delete facts but do not engender wholesale revision of perspective.

an empirical question just how often hardware biases the definition of a cognitive problem. *A priori* one would expect a continuum of problems from the most situated—where the cognitive task cannot be correctly defined without a careful analysis of the possible compliances and possible agent environment invariants—to highly abstract problems, such as word problems, number problems, puzzles and so forth, where the task is essentially abstract, and its implementation in the world is largely irrelevant to performance.[6]

Ultimately, Brooks' rejection of disembodied AI is an empirical challenge: for a large class of problems facing an acting creature the only reliable method of discovering how they can succeed, and hence what their true cognitive skills are, is to study them *in situ*.

Frequently this is the way of foundational questions. One theorist argues that many of the assumptions underpinning the prevailing methodology are false. He then proposes a new methodology and looks for empirical support.

But occasionally it is possible to offer, in addition to empirical support, a set of purely philosophical arguments against a methodology.

3.1. Philosophical objections to disembodied AI

At the top level we may distinguish two philosophical objections: first, that knowledge level accounts which leave out a theory of the body are too incomplete to serve the purpose for which they were proposed. Second, that axiomatic knowledge accounts fail to capture all the knowledge an agent has about a domain. Let us consider each in turn.

3.1.1. Why we need a theory of the body

The adequacy of a theory, whether in physics or AI, depends on the purpose it is meant to serve. It is possible to identify three rather different purposes AI theorists have in mind when they postulate a formal theory of the commonsense world. An axiomatic theory T of domain D is:

(1) adequate for *robotics* if it can be used by an acting perceiving machine to achieve its goals when operating in D;

(2) adequate for a *disembodied rational planner* if it entails all and only the intuitive truths of D as expressed in the language of the user of the planner;

(3) adequate for *cognitive science* if it effectively captures the knowledge of D which actual agents have.

[6] Clearly there are limits to how deviantly an abstract task may be implemented without effecting performance. Isomorphs of tic-tac-toe and the Tower of Hanoi are notoriously more difficult to solve than the standard problems. But the success in solving a problem often depends on finding its abstract structure—on understanding the constraints and options. Particular implementations or encodings of problems may make discovering this structure especially hard. But whenever success crucially depends on being mindful of that structure, knowledge level accounts of the problem are particularly appropriate.

The philosophical arguments I will now present are meant to show that a formal theory of D, unless accompanied by a theory about the sensori-motor capacities of the creature using the theory, will fail no matter which purpose a theorist has in mind. Theories of conceptualizations alone are inadequate, they require theories of embodiment.

Inadequacy for robotics According to Nilsson, the touchstone of adequacy of a logicist theory is that it marks the necessary domain distinctions and makes the necessary domain predictions for an acting perceiving machine to achieve its goals. Theoretical adequacy is a function of four variables: D: the actual subject-independent properties of a domain; P: the creature's perceptual capacities; A: the creature's action repertoire; and G: the creature's goals. In principle a change in any one of these can affect the theoretical adequacy of an axiomatization. For changes in perceptual abilities, no less than changes in action abilities or goals may render domain distinctions worthless, invisible to a creature.

If axioms are adequate only relative to $(D\ P\ A\ G)$ then formal theories are strictly speaking untestable without an account of $(D\ P\ A\ G)$. We can never know whether a given axiom set captures the distinctions and relations which a particular robot will need for coping with D. We cannot just assume that T is adequate if it satisfies our own intuitions of the useful distinctions inherent in a domain. The intuitions we ourselves have about the domain will be relative to our own action repertoire, perceptual capacities, and goals. Nor will appeal to model theory help. Model theoretic interpretations only establish consistency. They say nothing about the appropriateness, truth or utility of axiom sets for a given creature.

Moreover, this need to explicitly state A, P, and G is not restricted to robots or creatures having substantially different perceptual-motor capacities to our own. There is always the danger that between any two humans there are substantive differences about the intuitively useful distinctions inherent in a domain. The chemist, for instance, who wishes to axiomatize the knowledge a robot needs to cope with the many liquids it may encounter, has by dint of study refined his observational capacities to the point where he or she can notice theoretical properties of the liquid which remain invisible to the rest of us. She will use in her axiomatizations primitive terms that she believes are observational. For most of us they are not. We require axiomatic connections to tie those terms to more directly observational ones. As a result, there is in all probability a continuum of formal theories of the commonsense world ranging from ones understandable by novices to those understandable only by experts. Without an account of the observational capacities presupposed by a theory, however, it is an open question just which level of expertise a given T represents.

It may be objected that an account of the observational capacities presupposed by a theory is not actually part of the theory but of the metatheory of use—the theory that explains how to *apply* the theory. But this difference is in name alone. The domain knowledge that is required to tie a predicate to the observational conditions that are relevant to it is itself substantial. If a novice is to use the expert's theory he will have to know how to make all things considered judgements about whether a given phenomenon is an A-type event or B-type event. Similarly if the expert is to use the novice's theory he must likewise consult the novice's theory to decide the best way to collapse observational distinctions he notices. In either case, it is arbitrary where we say these world linking axioms are to be found. They are part and partial of domain knowledge. But they form the basis for a theory of embodiment.

Inadequacy for disembodied rational planners Despite the generality of the argument above it is hard to reject the seductive image of an omniscient angel—a disembodied intellect who by definition is unable to see or act—who nonetheless is fully knowledgeable of the properties of a domain and is able to draw inferences, make predictions and offer explanations in response to questions put to it.

The flaw in this image of a disembodied rational planner, once again, is to be found in the assumption that we can make sense of the angel's theoretical language without knowing how it would be hooked up to a body with sensors and effectors. Without some idea of what a creature would perceive the best we can do to identify the meaning it assigns to terms in its theory is to adopt a model theoretic stance and assume the creature operates with a consistent theory. In that case, the semantic content of a theory will be exhausted by the set of models satisfying it. Naturally, we would like to be able to single out one model, or one model family, as the *intended* models—the interpretation the angel has in mind when thinking about that theory. But there is no principle within model theory which justifies singling out one model as the intended model. Without some further ground for supposing the angel has one particular interpretation in mind we must acknowledge that the reference of the expressions in its theories are inscrutable.

It is not a weakness of model theory that it fails to state what a user of a language thinks his expressions are *about*. Model theory is a theory of validity, a theory of logical consequence. It states conditions under which an axiom set is consistent. It doesn't purport to be a theory of intentionality or a theory of meaning. This becomes important because unless all models are isomorphic to the intended model there will be possible interpretations that are so ridiculous given what we know that the axiom set is obviously empirically false. We know it doesn't correctly describe the entities and relations of the domain in question.

The way out of the model-theoretic straightjacket is once again by means of translation axioms linking terms in the axiom set to terms in our ordinary language. Thus if the angel uses a term such as "supports" as in "if you move a block supporting another block, the supported block moves" we assume that the meaning the angel has in mind for *support* is the same as that which we would have in the comparable English sentence. But now a problem arises. For unless we specify the meaning of these terms in English we cannot be confident the angel's theory is empirically adequate. The reason we must go this extra yard is that there are still too many possible interpretations of the terms in the axiom set. For instance, does the axiom "if you move a block supporting another, the supported block moves" seem correct? Perhaps. But consider cases where the upper block is resting on several lower blocks each supporting a corner of the upper block. Any single lower block can now be removed without disturbing the upper. Hence the axiom fails.

Were these cases intended? Exactly what range of cases did the angel have in mind? Without an account of intentionality, an account which explains what the angel would be disposed to recognize as a natural case and what as a deviant case, we know too little about the meaning of the angel's axioms to put them to use. Translation into English only shifts the burden because we still need to know what an English speaker would be disposed to recognize as a natural case and what as a deviant case. Without a theory of embodiment these questions are not meaningful.

Inadequacy for cognitive science I have been arguing that axiomatic accounts of common sense domains are incomplete for both robots and angels unless they include axioms specifying sensori-motor capacities, dispositions, and possibly goals. For the purposes of cognitive science, however, we may add yet another requirement to this list: that the predicates appearing in the axioms be extendable to new contexts in roughly the way the agents being modelled extend their predicates. We cannot say we have successfully captured the knowledge a given agent has about a domain unless we understand the concepts (or recognitional dispositions) it uses.

For instance, suppose an axiomatization of our knowledge of the blocks world fails to accommodate our judgements about novel blocks world cases. This will occur, for example, if we try to use our axioms of cubic blocks worlds to apply to blocks worlds containing pyramids. When our cubic blocks world axiomatization generates false predictions of this broader domain, shall we say the axiomatization fails to capture the single conceptualization of both worlds we operate with? Or shall we rather say that we must operate with more than one set of blocks world conceptions—one apt for cubic blocks, another for pyramidal, and so forth? One major school of thought maintains that it is the nature of human concepts that they be extendable to new domains without wholesale overhauling [19, 20]. Indeed that virtually all concepts, it is suggested, have this extensibility property.

Yet if extensibility is a feature of our conceptualizations then no axiomatization of our knowledge will be psychologically correct unless it also includes a set of axioms or principles for determining how we will extend our concepts to new domains. Axiomatizations without these principles will be too static, regularly giving rise to false predictions. On the other hand, extensibility dispositions cannot be stated without making reference to our sensori-motor dispositions and goals. Since these cannot be given without a theory of the agent's sense organs etc, axiomatizations in cognitive science must include a theory of embodiment.

3.1.2. Essential indexicality

The second set of arguments to show that an axiomatic theory of commonsense domains fail to capture all the knowledge the agents have about those domains turns on the rather severe assumptions implicit in model-theoretic interpretations of axioms that it be possible to state the intended interpretation of an axiom set in the language of sets and properties of objective spatial temporal regions. If it can be shown that systems often think about the world indexically, in an *egocentric* fashion, which cannot be adequately interpreted in terms of properties of objective space time regions, then there is some knowledge that an axiomatic theory fails to capture.

For example, my knowledge that my eyeglasses are *over there*, on my right, is not properly captured by describing my relation to a set of objective spatio-temporal models or geometric structures, because *over there* is not a standard function from words to worlds. If I am working with a data glove and manipulating objects on a display screen, *over there* means somewhere in data glove space. Similarly, if I am looking through a telescope, or I am wearing vision distorting glasses, what I mean when I say *over there* is not something context-independent; it very much matters on my action and perception space. What my knowledge of *over there* consists in is a set of dispositions to orient myself, to take certain actions which presuppose the location of the object relative to the type of actions I might perform. These dispositions cannot be described in terms of the public world of space and time, however, because they may have nothing to do with that shared world.[7]

Now if microtheories are meant to explain what we know about a domain that permits us to perform rational actions in that domain—for instance, if the microtheory of liquids is to partly explain why I open the tops of bottles, and upend them to extract their liquid contents—then that microtheory presupposes that we have the concept of *upending*. Yet if *upending* is a term that is meaningful egocentrically—and it must be for I may upend a bottle in data glove space—then our liquid microtheory does not capture our conceptual knowledge correctly. Many of the concepts we have are grounded in our egocentric understanding of our world of action and perception. Logicists tend

[7] The position I am cursorily describing derives from Gareth Evans in lecture and in [8].

to treat all concepts as designating entities in the public domain.[8] It is possible to introduce new constructs, such as perspectives, or situations to capture the agent's point of view on a space time region. But this still leaves unexplained the agent's perspective on virtual spaces which can be explained only by describing the agent's dispositions to behave in certain ways. Hence there are some things that an agent can know about a domain—such as where it is in a domain—which cannot be captured by standard axiomatic accounts.[9]

4. Is cognition rational kinematics?

I have been arguing that there are grave problems with the methodological assumption that cognitive skills can be studied in abstraction from the sensing and motor apparatus of the bodies that incorporate them. Both empirical and philosophical arguments can be presented to show that the body shows through. This does not vitiate the program of knowledge level theorists, but it does raise doubts about the probability of correctly modelling all cognitive skills on the knowledge-base/inference-engine model.

A further assumption related to disembodied AI is that we can use logic or English to track the trajectory of informational states a system creates as it processes a cognitive task. That is, either the predicate calculus or English can serve as a useful semantics for tracking the type of computation that goes on in cognition. They are helpful metalanguages.

From the logicist's point of view, when an agent computes its next behaviour it creates a trajectory of informational states that are *about* the objects, functions and relations designated in the designer's conceptualization of the environment. This language is, of course, a logical language. Hence the transitions between these informational states can be described as *rational transitions* or inferences in that logical language. If English is the semantic metalanguage, then rational transitions between sentences will be less well-defined, but ought nonetheless to make sense as *reasonable*.

There are two defects with this approach. First, that it is parochial: that in fact there are many types of computation which are not amenable to characterization in a logical metalanguage, but which still count as cognition. Second, because it is easy for a designer to mistake his own conceptualization for a machine's conceptualization there is a tendency to misinterpret the machine's informational trajectory, often attributing to the machine a deeper grasp of the world than is proper.

[8] For a brief account of the advantages of conceiving of the world as a public space, see my commentary on Rod Brooks [16].

[9] A third argument against model theoretic interpretations of knowledge is *inconsistency*. If there is an inconsistency in what I know about liquids, then there can be no models of this knowledge set. So I must know nothing at all. But of course I do know much about liquids, I just happen to be mistaken in one of my beliefs. Efforts to deal with such inconsistency exist in the literature [2].

Argument 1. Consider the second objection first. As mentioned earlier, it is necessary to distinguish those cases where:

(1) the designer uses concepts to describe the environment which the machine does not understand and perhaps could not;
(2) the designer uses only those concepts which the machine grasps, but the two represent those concepts differently;
(3) both designer and machine use the same concepts and encode them in the same way.

The first two cases concern the appropriate metalanguage of design, the last the object language of processing. Our goal as scientists is to represent a creature's cognition as accurately as possible, both so we can verify what it is doing, hence debug it better, and so we can design it better from the outset.

The trouble that regularly arises, though, is that the designer has a conceptualization of the task environment that is quite distinct from that of the system. There is always more than one way of *specifying* an ability, and more than one way of specifying an environment of action. Choice of a metalanguage should be made on pragmatic grounds: which formalism most simplifies the designer's task? But lurking in the background is the worry that if the designer uses a metalanguage that invokes concepts the system simply does not or could not have, then he may propose mistaken designs which he later verifies as correct using the same incorrect metalanguage.

For example, suppose we wish to design a procedure controlling a manipulator able to draw a circle using a pair of compasses. In our conceptualization we talk of a locus of points equidistant from a third point. Does the system itself operate with that conceptualization? Does it have *implicit* concepts of *locus*, *equidistance* and *points*?

Why does it matter? Well, suppose we now have the manipulator attempt to draw a circle on a crumpled piece of paper. The naive procedure will not produce a curve whose distance on the crumpled surface is equidistant. Its design works for flat surfaces, not for arbitrary surfaces. Yet if a system did have concepts for equidistance, locus and points it ought to be *adaptive* enough to accommodate deformations in surface topology. To be sure such a machine would have to have some way of sensing topology. That by itself is not enough, though. It is its dispositions to behave in possible worlds that matters. This is shown by the old comment that whether I have the concept *chordate* (creature having a heart) or *renate* (creature having kidneys) cannot be determined by studying my normal behaviour alone [34]. In normal worlds, all chordates are renates. Only in counterfactual worlds—where it is possible to come across viable creatures with hearts but no kidneys—could we display our individuating dispositions. The upshot is that a designer cannot assume that his characterization of the informational trajectory of a creature is correct, unless he confirms certain claims about the creature's dispositions to behave in a range of further contexts. Sometimes these contexts lie outside the narrow task he is building a cognitive skill for.

None of the above establishes that English is inadequate. It just shows that it is easy to make false attributions of content. The criticism that logic and natural language are not adequate metalanguages arises as soon as we ask whether they are expressive enough to describe some of the bizarre concepts systems with funny dispositions will have. In principle, both logic and English are expressive enough to capture any comprehensible concept. But the resulting characterization may be so long and confusing that it will be virtually incomprehensible. For instance, if we try to identify what I have been calling the implicit concepts of the compass controller we will be stymied. If the system could talk what would it say to the question: Can a *circle* be drawn in a space measured with a non-Euclidian metric? What nascent idea of equidistance does it have? Its inferences would be so idiosyncratic that finding an English sentence or reasonable axiomatic account would be out of the question. English and logic are the wrong metalanguages to characterize such informational states.

What is needed is more in the spirit of a functional account of informational content [1]. Such semantics are usually ugly. For in stating the role an informational state plays in a system's dispositions to behave we characteristically need to mention myriad other states, since the contribution of a state is a function of other states as well.

Accordingly, not all informational states are best viewed as akin to English sentences. If we want to understand the full range of cognitive skills—especially those modular ones which are not directly hooked up to central inference—we will need to invoke some other language for describing information content. Frequently the best way to track a computation is not as a rational trajectory in a logical language.

Argument 2. The need for new languages to describe informational content has recently been re-iterated by certain connectionists who see in parallel distributing processing a different style of computation. Hewitt and Gasser have also emphasized a similar need for an alternative understanding of the computational processes occurring in distributed AI systems. It is old fashioned and parochial to hope for a logic-based denotational semantics for such systems.

The PDP concern can be stated as follows: in PDP computation vectors of activation propagate through a partially connected network. According to Smolensky [41] it is constructive to describe the behaviour of the system as a path in tensor space. The problem of interpretation is to characterize the significant events on this path. It would be pleasing if we could say "now the network is extracting the information that p, now the information that q", and so on, until the system delivers its answer. Unfortunately, though, except for input and output vectors—whose interpretation we specifically set—the majority of vectors are not interpretable as carrying information which can be easily stated in English or logic. There need be no one–one mapping between significant events in the system's tensor space trajectory and its path in propositional space. Smolensky—whose argument this is—suggests that much of this intermediate processing is interpretable at the subconceptual level where the basic elements of meaning differ from those we have words for in English.[10]

In like manner, Hewitt and Gasser offer another argument for questioning whether we can track the information flowing through a complex system in propositional form. The question they ask is: How are we to understand the content of a message sent between two agents who are part of a much larger matrix of communicating agents. Superficially, each agent has its own limited perspective on the task. From agent-1's point of view, agent-2 is saying p, from agent-3's point of view, agent-2 is saying q. Is there a right answer? Is there a God's eye perspective that identifies the true content and gives the relativized perspective of each agent? If so, how is this relativized meaning to be determined? We will have to know not only whom the message is addressed to, but what the addressee is expecting, and what it can *do* with the message. Again, though, once we focus on the effects which messages have on a system we leave the simple world of denotational semantics and opt for functional semantics. Just how we characterize *possible effects*, however, is very different than giving a translation of the message in English. We will need a language for describing the behavioural dispositions of agents.

Cognition as rational inference looks less universal once we leave the domain of familiar sequential processing and consider massively parallel architectures.

5. Can cognition be studied separately from learning?

In a pure top-down approach, we assume it is possible to state what a system knows without stating how it came to that knowledge. The two questions, competence and acquisition can be separated. Learning, on this view, is a switch that can be turned on or off. It is a box that takes an early conceptualization and returns a more mature conceptualization. Thus learning and con-

[10] One way of seeing the problem is to recognize that in a simple feed-forward network a given hidden unit can be correlated with a (possibly nested) disjunction of conjunctions of probabilities of input features. A vector, therefore, can be interpeted as a combination of these. The result is a compound that may make very little sense to us. For instance, it might correspond to a distribution over the entire feature set. Thus a single node might be tuned to respond to the weighted conjunction of features comprising the tip of my nose, my heel, plus the luminesence of my hands, or the weighted conjunction of. . . . Moreover, if we do not believe that the semantics of networks is correlational but rather functional we will prefer to interpret the meaning of a node to be its contribution (in conjunction with its superior nodes) to the capacity to classify.

ceptualization are sufficiently distinct that the two can be studied separately. Indeed, learning is often understood as the mechanism for generating a trajectory of conceptualizations. This is clearly the belief of logic theorists and developmental psychologists who maintain that what an agent knows at a given stage of development is a theory, not fundamentally different in spirit than a scientific theory, about the domain [4].

There are several problems with this view. First, it assumes we can characterize the instantaneous conceptualization of a system without having to study its various earlier conceptualizations. But what if we cannot *elicit* the system's conceptualization using the standard techniques? To determine what a competent PDP system, for example, would know about its environment of action, it is necessary to train it until it satisfies some adequacy metric. We cannot say in advance what the system will know if it is perfectly competent because there are very many paths to competence, each of which potentially culminates in a different solution. Moreover if the account of PDP offered above is correct it may be impossible to characterize the system's conceptualization in a logical language or in English. It is necessary to analyze its dispositions. But to do that one needs an actual implementation displaying the competence. Hence the only way to know what a PDP system will know if it is competent is to build one and study it. A purely top-down stance, which asssumes that learning is irrelevant, is bound to fail in the case of PDP.

A second argument against detaching knowledge and learning also focusses on the *in practice* unpredictable nature of the learning trajectory. In Soar it is frequently said that chunking is more than mere speedup [35]. The results of repeatedly chunking solutions to impasses has a nonlinear effect on performance. Once we have nonlinear effects, however, we cannot predict the evolution of a system short of running it. Thus in order to determine the steady state knowledge underpinning a skill we need to run Soar with its chunking module on.[11]

A final reason we cannot study what a system knows without studying how it acquires that knowledge is that a system may have been special design features that let it acquire knowledge. It is organized to self-modify. Hence we cannot predict what knowledge it may contain unless we know how it integrates new information with old. There are many ways to self-modify.

For instance, according to Roger Schank, much of the knowledge a system contains is lodged in its indexing scheme [41]. As systems grow in size they generally have to revise their indexing scheme. The results of this process of revision cannot be anticipated *a priori* unless we have a good idea of the earlier indexing schemes. The reason is that much of its knowledge is stored in cases. Case knowledge may be sensitive to the order the cases were encountered.

[11] We can, of course, hand-simulate running the system and so predict its final states. But I take it this is not a significant difference from running Soar itself.

Consequently, we can never determine the knowledge a competent system has unless we know something of the cases it was exposed to and the order they were met. History counts.

This emphasis on cases goes along with a view that much of reasoning involves noticing analogies to past experiences. A common corrolary to this position is that concepts are not context-free intensions; they have a certain open texture, making it possible to flexibly extend their use and to apply them to new situations in creative ways. An agent which understands a concept should be able to recognize and generate analogical extensions of its concepts to new contexts.

Once we view concepts to be open textured, however, it becomes plausible to suppose that a concept's meaning is a function of history. It is easier to see an analogical extension of a word if it has already been extended in that direction before. But then, we can't say what an agent's concept of "container" is unless we know the variety of contexts it has seen the word in. If that is so, it is impossible to understand a creature's conceptualization in abstraction from its learning history. Much of cognition cannot be studied independently of learning.

6. Is the architecture of cognition homogeneous?

The final issue I will discuss is the claim made by Newell et al. that cognition is basically the product of running programs in a single architecture. According to Newell, too much of the research in AI and cognitive science aims at creating independent representational and control mechanisms for solving particular cognitive tasks. Each investigator has his or her preferred computational models which, clever as they may be, rarely meet a further constraint that they be integratable into a unified account of cognition. For Newell

> Psychology has arrived at the possibility of unified theories of cognition—theories that gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition [30].

The idea that there might be a general theory of intelligence is not new. At an abstract level anyone who believes that domain knowledge plus inferential abilities are responsible for intelligent performance, at least in one sense, operates with a general theory of cognition. For, on that view, it is knowledge, ultimately, that is the critical element in cognition.

But Newell's claim is more concrete: not only is knowledge the basis for intelligence; knowledge, he argues further, will be encoded in a Soar-like mechanism. This claim goes well beyond what most logicists would maintain. It is perfectly consistent with logicism that knowledge may be encoded, implemented or embedded in any of dozens of ways. A bare commitment to specification of cognitive skills at the knowledge level is hardly grounds for expecting a small set of "underlying mechanisms, whose interactions and compositions provide the answers to all the questions we have—predictions, explanations, designs, controls" [30, p. 14] pertaining to the full range of cognitive performances. The Soar project, however, is predicated on this very possibility. The goal of the group is to test the very strong claim that underpinning problem solving, decision making, routine action, memory, learning, skill, even perception and motor behaviour, there is a single architecture "a single system [that] produces all aspects of behaviour . . . Even if the mind has parts, modules, components, or whatever, they mesh together . . ." and work in accordance with a small set of principles.

It is not my intent to provide serious arguments for or against this position. I mention it largely because it is such a deep committment of the Soar research program and therefore an assumption that separates research orientations. The strongest support for it must surely be empirical, and it will become convincing only as the body of evidence builds up. There can be little doubt, though, that it is an assumption not universally shared.

Minsky, for instance, in *Society of Mind* [28], has argued that intelligence is the product of hundreds, probably thousands of specialized computational mechanisms he terms agents. There is no homogenous underlying architecture. In the society of mind theory, mental activity is the product of many agents of varying complexity interacting in hundreds of ways. The very purpose of the theory is to display the variety of mechanisms that are likely to be useful in a mind-like system, and to advocate the need for diversity. Evolution, Minsky, emphasizes is an opportunistic tinkerer likely to co-opt existing mechanisms in an *ad hoc* manner to create new functions meeting new needs. With such diversity and ad hoccery it would be surprising if most cognitive performances were the result of a few mechanisms comprising a principled architecture.

Brooks in a similar manner sets out to recreate intelligent capacities by building layer upon layer of mechanism, each with hooks into lower layers to suppress or bias input and output. Again, no non-empirical arguments may be offered to convince skeptics of the correctness of this view. The best that has been offered is that the brain seems to have diverse mechanisms of behaviour control, so it is plausible that systems with comparable functionality will too.

Again there is no quick way to justify the assumption of architecture homogeneity. More than any other foundational issue this is one for which non-empirical or philosophical arguments are misplaced.

7. Conclusion

I have presented five dimensions—five big issues—which theorists in AI, either tacitly or explicitly, take a stand on. Any selection of issues is bound to have a personal element to them. In my case I have focussed most deeply on the challenges of embodiment. How reliable can theories of cognition be if they assume that systems can be studied abstractly, without serious concern for the mechanisms that ground a system's conceptualization in perception and action? But other more traditional issues are of equal interest. How central is the role which knowledge plays in cognitive skills? Can most of cognition be seen as inference? What part does learning or psychological development play in the study of reasoning and performance? Will a few mechanisms of control and representation suffice for general intelligence? None of the arguments presented here even begin to be decisive. Nor were they meant to be. Their function is to encourage informed debate of the paramount issues informing our field.

Acknowledgement

I thank Farrel Ackerman, John Batali, Danny Bobrow, Pat Hayes, Paul Kube, Brian Smith and Patrick Winston for helpful and fun conversations on the topics of this paper.

References

[1] L. Birnbaum, Rigor mortis: a response to Nilsson's "Logic and artificial intelligence", *Artif. Intell.* **47** (1991) 57–77, this volume.

[2] M. Brandon and N. Rescher, *The Logic of Inconsistency* (Basil Blackwell, Oxford, 1978).

[3] R.A. Brooks, Intelligence without representation, *Artif. Intell.* **47** (1991) 139–159, this volume.

[4] S. Carey, *Conceptual Change in Childhood* (MIT Press/Bradford Books, Cambridge, MA, 1985).

[5] N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA, 1965).

[6] N. Chomsky, *Knowledge of Language: Its Nature Origin and Use* (Preager, New York, 1986).

[7] A. Cussins, Connectionist construction of concepts, in: M. Boden, ed., *Philosophy of Artificial Intelligence* (Oxford University Press, Oxford, 1986).

[8] G. Evans, *Varieties of Reference* (Oxford University Press, Oxford, 1983).

[9] J.A. Fodor, *Language of Thought* (Harvard University Press, Cambridge, MA, 1975).

[10] J.A. Fodor, *Psychosemantics* (MIT Press, Cambridge, MA, 1987).

[11] J.A. Fodor and Z.W. Pylyshyn, Connectionism and cognitive architecture: a critical analysis, *Cognition* **28** (1988) 3–71.

[12] L. Gasser, Social conceptions of knowledge and action: DAI foundations and open systems semantics, *Artif. Intell.* **47** (1991) 107–138, this volume.

[13] P.J. Hayes, A critique of pure treason, *Comput. Intell.* **3** (3) (1987).

[14] C. Hewitt, Open Information Systems Semantics for Distributed Artificial Intelligence, *Artif. Intell.* **47** (1991) 79–106, this volume.

[15] J.R. Hobbs and R. Moore, eds., *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985).

[16] D. Kirsh, Today the earwig, tomorrow man?, *Artif. Intell.* **47** (1991) 161–184, this volume.

[17] K. Konolige, Belief and incompleteness, in: J.R. Hobbs and R. Moore, eds., *Formal Theories of the Commonsense World* (Ablex, Norwood, NJ, 1985).

[18] T. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, IL, 1962).

[19] G. Lakoff, *Women, Fire, Dangerous Things: What Categories Reveal about the Mind* (University of Chicago Press, Chicago, IL, 1987).

[20] R. Langacker, *Foundations of Cognitive Grammar* (Stanford University Press, Stanford, CA, 1987).

[21] D.B. Lenat and R.V. Guha, *Building Large Knowledge-Based Systems, Representation and Inference in the Cyc Project* (Addison-Wesley, Reading, MA, 1989).

[22] D.B. Lenat and J.S. Brown, Why AM and EURISKO appear to work, *Artif. Intell.* **23** (1984) 269–294.

[23] D.B. Lenat and E.A. Feigenbaum, On the thresholds of knowledge, *Artif. Intell.* **47** (1991) 185–250, this volume.

[24] H.J. Levesque, Knowledge representation and reasoning, in: *Annual Review of Computer Science* **1** (Annual Reviews Inc., Palo Alto, CA, 1986) 255–287.

[25] J. Mandler, How to build a baby 2, unpublished manuscript.

[26] D. Marr, *Vision* (Freeman, San Francisco, CA, 1982).

[27] J.L. McClelland, D.E. Rumelhart and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **2**: *Psychological and Biological Models* (MIT Press/Bradford Books, Cambridge, MA, 1986).

[28] M.L. Minsky, *The Society of Mind* (Simon and Schuster, New York, 1986).

[29] R. Montague, *Formal Philosophy: Selected Papers of Richard Montague*, edited by R.H. Thomason (Yale University Press, New Haven, CT, 1974).

[30] A. Newell, Unified theories of cognition: the William James lectures, manuscript.

[31] A. Newell, P.S. Rosenbloom and J.E. Laird, Symbolic architectures for cognition, in: M. Posner, ed., *Foundations of Cognitive Science* (MIT Press, Cambridge, MA, 1989).

[32] A. Newell and H.A. Simon, *Human Problem Solving* (Prentice-Hall, Englewood Cliffs, NJ, 1972).

[33] N.J. Nilsson, Logic and artificial intelligence, *Artif. Intell.* **47** (1991) 31–56, this volume.

[34] W.V.O. Quine, *Word and Object* (MIT Press, Cambridge, MA, 1960).

[35] P.S. Rosenbloom, J.E. Laird, A. Newell and R. McCarl, A preliminary analysis of the Soar architecture as a basis for general intelligence, *Artif. Intell.* **47** (1991) 289–325, this volume.

[36] S.J. Rosenschein, The logicist conception of knowledge is too narrow—but so is McDermott's, *Comput. Intell.* **3** (3) (1987).

[37] S.J. Rosenschein and L.P. Kaebling, The synthesis of machines with provably epistemic properties, in: J.Y. Halpern, ed., *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, Los Altos, CA, 1986) 83–98.

[38] D.E. Rumelhart, J.L. McClelland and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**: *Foundations* (MIT Press/ Bradford Books, Cambridge, MA, 1986).

[39] D.E. Rumelhart et al., Schemata and sequential thought processes in PDP models, in: J.L. McClelland, D.E. Rumelhart and the PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **2**: *Psychological and Biological Models* (MIT Press/Bradford Books, Cambridge, MA, 1986).

[40] R.C. Schank, *Dynamic Memory* (Erlbaum, Hillsdale, NJ, 1985).

[41] R.C. Schank and C. Riesbeck, *Inside Computer Understanding* (Erlbaum, Hillsdale, NJ, 1981).

[42] B.C. Smith, The owl and the electric encyclopedia, *Artif. Intell.* **47** (1991) 251–288, this volume.

[43] P. Smolensky, On the proper treatment of connectionism, *Behav. Brain Sci.* **11** (1988) 1–23.