FOUNDATIONS OF ARTIFICIAL INTELLIGENCE

EDITED BY DAVID KIRSH

MIT/ELSEVIER

Foundations of Artificial Intelligence

edited by

David Kirsh

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

First MIT Press edition, 1992

Reprinted from *Artificial Intelligence: An International Journal*, Volume 47, Numbers 1–3, 1991. The MIT Press has exclusive license to sell this English-language book edition throughout the world.

Printed and bound in the Netherlands. This book is printed on acid-free paper.

Library of Congress Cataloging-in-Publication Data

Foundations of artificial intelligence / edited by David Kirsh. — 1st MIT Press ed.
p. cm. — (Special issues of Artificial intelligence, an international journal)
"A Bradford Book."
"Reprinted from Artificial intelligence, an international journal, volume 47, numbers 1–3, 1991"—T.p. verso.
Includes bibliographical information and index.
ISBN 0-262-61075-2
1. Artificial intelligence. I. Kirsh, David. II. Series.
Q335.5.F68 1992
006.3—dc20 91-18104
CIP

Contents

D. KIRSH
1 Foreword

D. KIRSH
3 Foundations of AI: the big issues

N.J. NILSSON
31 Logic and artificial intelligence

L. BIRNBAUM
57 Rigor mortis: a response to Nilsson's "Logic and artificial intelligence"

C. HEWITT
79 Open Information Systems Semantics for Distributed Artificial Intelligence

L. GASSER
107 Social conceptions of knowledge and action: DAI foundations and open systems semantics

R.A. BROOKS
139 Intelligence without representation

D. KIRSH
161 Today the earwig, tomorrow man?

D.B. LENAT AND E.A. FEIGENBAUM
185 On the thresholds of knowledge

B.C. SMITH
251 The owl and the electric encyclopedia

P.S. ROSENBLOOM, J.E. LAIRD, A. NEWELL AND R. MCCARL
289 A preliminary analysis of the Soar architecture as a basis for general intelligence

D.A. NORMAN
327 Approaches to the study of intelligence

347 INDEX

Artificial Intelligence 47 (1991) 1
Elsevier

Foreword

The intent of this special issue on the Foundations of AI is to critically evaluate the fundamental assumptions underpinning the dominant approaches to AI. Theorists historically associated with each position were originally invited to a workshop at Endicott House, MIT, in 1987. They were asked to write a paper identifying the basic tenets of their position, to discuss the principles underpinning the method or approach, to describe the natural type of problems and tasks in which the approach succeeds, to explain where the power resides in the method or approach, and then to discuss its scope and limits. Theorists generally skeptical of the position were similarly asked to evaluate the source of power of the method or approach and to state why they thought the method/approach works and why it fails. Discussions and presentations there formed the basis for the papers published here. We gratefully acknowledge the support for the workshop provided for by the MIT Artificial Intelligence Laboratory, National Science Foundation, and AAAI.

David Kirsh
Department of Cognitive Science
University of California, San Diego
La Jolla, CA, USA

0004-3702/91/$03.50

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.

0004-3702/91/$03.50

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].

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 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

- 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.

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 axiomatizato 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].

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 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).

Artificial Intelligence 47 (1991) 31–56
Elsevier

Logic and artificial intelligence

Nils J. Nilsson
Computer Science Department, Stanford University, Stanford, CA 94305, USA

Received February 1989

Abstract

Nilsson, N.J., Logic and artificial intelligence, Artificial Intelligence 47 (1990) 31–56.

The theoretical foundations of the logical approach to artificial intelligence are presented. Logical languages are widely used for expressing the declarative knowledge needed in artificial intelligence systems. Symbolic logic also provides a clear semantics for knowledge representation languages and a methodology for analyzing and comparing deductive inference techniques. Several observations gained from experience with the approach are discussed. Finally, we confront some challenging problems for artificial intelligence and describe what is being done in an attempt to solve them.

1. Introduction

Until a technological endeavor achieves a substantial number of its goals, several competing approaches are likely to be pursued. So it is with artificial intelligence (AI). AI researchers have programmed a number of demonstration systems that exhibit a fair degree of intelligence in limited domains (and some systems that even have commercial value). However, we are still far from achieving the versatile cognitive skills of humans. And so research continues along a number of paths—each with its ardent proponents. Although successful AI systems of the future will probably draw upon a combination of techniques, it is useful to study the different approaches in their pure forms in order to highlight strengths and weaknesses. Here, I present my view of what constitutes the "logical approach" to AI.

Some of the criticisms of the use of logic in AI stem from confusion about what it is that "logicists" claim for their approach. As we shall see, logicism provides a point of view and principles for constructing languages and procedures used by intelligent machines. It certainly does not promise a ready-made apparatus whose handle needs only to be turned to emit intelligence. Indeed, some researchers who might not count themselves among those following a logical approach can arguably be identified with the logicist position. (See, for example, Smith's review of a paper by Lenat and Feigenbaum [28, 54].) Other,

0004-3702/91/$03.50

more naive, criticisms claim that since so much of human thought is "illogical" (creative, intuitive, etc.), machines based on logic will never achieve human-level cognitive abilities. But puns on the word "logic" are irrelevant for evaluating the use of logic in building intelligent machines; making "illogical" machines is no trouble at all!

In describing logic and AI, we first relate the logical approach to three theses about the role of knowledge in intelligent systems. Then we examine the theoretical foundations underlying the logical approach. Next, we consider some important observations gained from experience with the approach. Lastly, we confront some challenging problems for AI and describe what is being done in an attempt to solve them. For a textbook-length treatment of logic and AI see [12].

2. Artificial intelligence and declarative knowledge

The logical approach to AI is based on three theses:

Thesis 1. Intelligent machines will have knowledge of their environments.

Perhaps this statement is noncontroversial. It is probably definitional. Several authors have discussed what it might mean to ascribe knowledge to machines—even to simple machines such as thermostats [33, 48].

Thesis 2. The most versatile intelligent machines will represent much of their knowledge about their environments declaratively.

AI researchers attempt to distinguish between *declarative* and *procedural* knowledge and argue about the merits of each. (See, for example, [16, 60].) Roughly speaking, declarative knowledge is encoded explicitly in the machine in the form of sentences in some language, and procedural knowledge is manifested in programs in the machine. A more precise distinction would have to take into account some notion of *level* of knowledge. For example, a LISP program which is regarded as a program (at one level) is regarded (at a lower level) as a declarative structure that is interpreted by another program. Settling on precise definitions of procedural and declarative knowledge is beyond our scope here. Our thesis simply states that versatile intelligent machines will have (among other things) a place where information about the environment is stored explicitly in the form of sentences. Even though any knowledge that is ascribed to a machine (however represented in the machine) might be given a declarative interpretation by an outside observer, we will not say that the machine possesses declarative knowledge unless such knowledge is actually represented by explicit sentences in the memory of the machine.

When knowledge is represented as declarative sentences, the sentences are manipulated by reasoning processes when the machine is attempting to use that knowledge. Thus, the component that decides how to *use* declarative knowledge is separate from the knowledge itself. With procedural approaches to knowledge representation, knowledge use is inextricably intertwined with knowledge representation.

The first serious proposal for an intelligent system with declarative knowledge was by John McCarthy [32]. McCarthy noted the versatility of declaratively represented knowledge: it could be used by the machine even for purposes unforeseen by the machine's designer, it could more easily be modified than could knowledge embodied in programs, and it facilitated communication between the machine and other machines and humans. As he wrote later, "Sentences can be true in much wider contexts than specific programs can be useful" [36].

Smolensky [55] listed some similar advantages: "a. *Public access*: [Declarative] knowledge is accessible to many people; b. *Reliability*: Different people (or the same person at different times) can reliably check whether conclusions have been validly reached; c. *Formality*, *bootstrapping*, *universality*: The inferential operations require very little experience with the domain to which the symbols refer."

To exploit these advantages, the declaratively represented knowledge must, to a large extent, be *context free*. That is, the *meaning* of the sentences expressing the knowledge should depend on the sentences themselves and not on the external context in which the machine finds itself. The context-free requirement would rule out terms such as "here" and "now" whose meaning depends on context. Such terms are called *indexicals*.

Many database systems and expert systems can be said to use declarative knowledge, and the "frames" and "semantic networks" used by several AI programs can be regarded as sets of declarative sentences. On the other hand, there are several examples of systems that do not represent knowledge about the world as declarative sentences. Some of these are described in the other papers in this volume.

Thesis 3. For the most versatile machines, the language in which declarative knowledge is represented must be at least as expressive as first-order predicate calculus.

One might hope that a natural language such as English might serve as the language in which to represent knowledge for intelligent systems. If this were possible, then all of the knowledge already compiled in books would be immediately available for use by computers. Although humans somehow understand English well enough, it is too ambiguous a representational medium for present-day computers—the meanings of English sentences depend too much on the contexts in which they are uttered and understood.

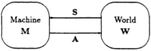
AI researchers have experimented with a wide variety of languages in which to represent sentences. Some of these languages have limited expressive power. They might not have a means for saying that one or another of two facts is true without saying which fact is true. Some cannot say that a fact is not true without saying what is true instead. They might not be able to say that *all* the members of a class have a certain property without explicitly listing each of them. Finally, some are not able to state that at least one member of a class has a certain property without stating which member does. First-order predicate calculus, through its ability to formulate disjunctions, negations, and universally and existentially quantified sentences, does not suffer from these limitations and thus meets our minimal representational requirements.

3. Foundations of the logical approach

In addition to the three theses just stated, the logical approach to AI also embraces a point of view about what knowledge is, what the world is, how a machine interacts with the world, and the role and extent of special procedures in the design of intelligent machines.

Those designers who would claim that their machines possess declarative knowledge about the world are obliged to say something about what that claim means. The fact that a machine's knowledge base has an expression in it like $(\forall x)\mathrm{Box}(x) \supset \mathrm{Green}(x)$, for example, doesn't by itself justify the claim that the machine *believes* all boxes are green. (The mnemonic relation constants that we use in our design aren't mnemonic for the machine! We could just as well have written $(\forall x)\mathrm{GO11}(x) \supset \mathrm{GO23}(x)$.)

There are different views of what it means for a machine possessing a database of sentences to believe the facts intended by those sentences. The view that I favor involves making some (perhaps unusual) metaphysical



$$see : \mathcal{W} \longrightarrow \mathcal{S}$$

$$mem : \mathcal{S} \times \mathcal{M} \longrightarrow \mathcal{M}$$

$$act : \mathcal{S} \times \mathcal{M} \longrightarrow \mathcal{A}$$

$$effect : \mathcal{A} \times \mathcal{W} \longrightarrow \mathcal{W}$$

Fig. 1. Machine and world.

a mathematical structure, but since our picture provides for the world to be affected by and affect itself and the intelligent machine, one shouldn't worry that our view of the world is impractically ethereal.)

Now, the designer of a machine that is to interact with the world never knows what the world objects, functions, and relations actually are. He must guess. Guessing involves *invention* on the designer's part. (Our machine designer is in the same predicament as is the scientist; scientists invent descriptions of the world and gradually refine them until they are more useful.) We use the term *conceptualization* to describe the designer's guess about the world objects, functions, and relations. The designer may not even be able to specify a single conceptualization; for example he may choose not to commit himself regarding whether an object he invents, say a block, has the color property green or blue. Thus, in general, the designer attempts to specify a set of conceptualizations such that, whatever the world actually is, he guesses it is a member of the set.

The designer realizes, of course, that his conceptualization might not accurately capture the world—even as he himself believes it to be. For example, his conceptualization may not discriminate between objects that he himself recognizes to be different but which can be considered to be the same considering his purposes for the machine. The designer need only invent a conceptualization that is *good enough*, and when and if it becomes apparent that it is deficient (and that this deficiency is the cause of inadequate machine performance), he can modify his conceptualization.

We stress that the objects guessed to exist in the world by the designer are *invented*. He is perfectly free to invent anything that makes the machine perform appropriately, and he doesn't ask whether or not some object *really* does or does not exist (whatever that might mean) apart from these invented structures. For many ordinary, concrete objects such as chairs, houses, people, and so on, we can be reasonably confident that our inventions mirror reality. But some of the things that we might want to include as world objects, such as *precambrian unconformities*, English sentences, *the Peloponnesian War*, π, and *truth*, have a somewhat more arbitrary ontological status. In fact, much of the designer's guess about the world may be quite arbitrary in the sense that other guesses would have suited his purposes equally well. (Even those researchers following other declarative, but putatively non-logical, approaches must invent the equivalent of objects, relations, and functions when they attempt to give their machines declarative knowledge.)

A logicist expresses his conceptualization of the world (for the machine) by a set of sentences. The sentences are made part of the machine's memory (comprising its state) and embody the machine's declarative knowledge. We assume that the sentences are in the first-order predicate calculus; this language and the sentences in it are constructed as follows: For every world object in the conceptualization we create an *object constant*; for every world relation, we the machine is attached to the world, as in Fig. 1, *mem* produces a sequence of states $\Delta_0, \Delta_1, \ldots, \Delta_i, \ldots$.

Even when the designer has a single intended interpretation in mind, Δ, in general, will be satisfied by a set of interpretations—the intended one among them. The designer must provide sufficient sentences in the knowledge base such that its models are limited—limited so that even though the set has more than one model, it doesn't matter given the purposes for the machine. (To the extent that it *does* matter, the designer must then provide more sentences.) In designing knowledge bases, it frequently happens that the designer's idea of the intended interpretation is changed and articulated by the very act of writing down (and reasoning with) the sentences.

So, a machine possessing a set of sentences *knows* about the world in the sense that these sentences admit of a set of models, and this set is the designer's best approximation to what the world actually is, given the purposes for the machine. The *actual* world might not even be in the set (the designer's guess might be wrong), so we really should be talking about the machine's *beliefs* rather than the machine's *knowledge*. But, following the tradition established by the phrase "knowledge-based systems," we will continue to speak of the machine's knowledge.

The machine's procedural knowledge is represented in the functions *mem* and *act*. The function *mem* changes the sentences and thereby changes the machine's state. Perhaps new sentences are added or existing ones are modified or deleted in response to new sensory information. The function *mem* may also produce a change in the machine's state in the absence of sensory information; changes to Δ may occur through processes of deduction or other types of inference as will be described below.

The machine's declarative knowledge affects its actions through the function *act*. We take *act* to be a function (over sets of sentences) that produces actions. Note that *act* can thus only respond to sentences *qua sentences*, that is, as strings of symbols. It is not a function of the models of these sentences!

Given this picture, we can identify a spectrum of design choices. At one end, *act* and *mem* are highly specialized to the tasks the machine is expected to perform and to the environment in which it operates. We might say, in this case, that the machine's knowledge is mainly *procedurally* represented. At the other extreme, *act* and *mem* are general purpose and largely independent of the application. All application-specific knowledge is represented in Δ. The machine's knowledge in this case can be said to be mainly *declaratively* represented. The logical approach usually involves a commitment to represent most of the machine's knowledge declaratively. For a proposal at the extreme declarative end, see [12, Chapter 13]. It is not yet known to what extent this goal can be achieved while maintaining reasonable efficiency.

Because the actions emitted by *act* depend on the syntactic form of the sentences in Δ, it is necessary for *mem* to be able to rewrite these sentences in the form appropriate to the task at hand. This aspect of *mem* we call *reasoning*. Imagine, for example, a robot designed to paint boxes green. Its sentence-to-action process, *act*, may include a *production rule* like "If Δ includes the sentence Box(η) for some value of η, paint the object denoted by η green." But suppose Δ includes the sentences ($\forall$x)Blue(x) $\supset$ Box(x) and Blue(G17) but not Box(G17) explicitly. We might expect that correct behavior for this robot would be to paint the object denoted by G17 green, but there is no sentence-to-action rule to accomplish that unless Box(G17) occurs explicitly in Δ. Constructing the sentence Box(G17) from the sentences ($\forall$x)Blue(x) $\supset$ Box (x) and Blue(G17) is an example of one kind of sentence manipulation, or *inference*, that we want *mem* to do.

Often, as in the box-painting example, the new sentence constructed from ones already in memory does not tell us anything new about the world. (All of the models of ($\forall$x)Blue(x) $\supset$ Box(x) and Blue(G17) are also models of Box(G17). Thus, adding Box(G17) to Δ does not reduce the set of models.) What the new sentence tells us was already implicitly said by the sentences from which it was constructed.

If all of the models of Δ are also models of a sentence ϕ, we say that Δ *logically entails* ϕ and write $\Delta \models \phi$. Among the computations that we might want *mem* to perform are those which add sentences to Δ that are logically entailed by Δ. One apparent problem in devising such computations is the prospect of having to check *all* the models of Δ to see if they are also models of ϕ. But, fortunately, there exist strictly syntactic operations on Δ that are able to compute logically entailed formulas.

We use the phrase *rule of inference* to refer to any computation on a set of sentences that produces new sentences. If ψ can be derived from Δ by a sequence of applications of rules of inference, we say that ψ can be *deduced* from Δ and write $\Delta \vdash \psi$. An example is the rule of inference called *modus ponens*. From any sentences of the form $\rho \supset \sigma$ and ρ, we can deduce the sentence σ by modus ponens. The process of logical deduction involves using a set of rules of inference to deduce additional sentences from a set of sentences. Interestingly, it happens that there are rules of inference, modus ponens is an example, that have the property that if $\Delta \vdash \phi$, then $\Delta \models \phi$. Such rules of inference are called *sound*.

Sound rules of inference are extremely important because they allow us to compute sentences that are logically entailed by a set of sentences using computations on the sentences themselves (and not on their models).

We can also find sets of inference rules that have the property that if $\Delta \models \phi$ then the rules (successively applied) will eventually produce such a ϕ. Such a set of inference rules is called *complete*.

Although all logicists typically incorporate sound inference rules as part of the calculations performed by *mem*, there is no necessary reason to limit *mem* to performing sound inferences. Other computations are often desirable. We will describe some of these later in the paper.

In summary, intelligent machines designed according to the logical approach are state-machines whose states are sets of sentences. Machine state transitions are governed by a function, *mem*, acting on the sentence sets and the inputs to the machine. An important, but not the only, component of *mem* is sound logical inference. Machine actions are governed by a function, *act*, of the machine's state and inputs. The intended interpretation of the sentences in a machine's state involves objects, functions, and relations that are the designer's guesses about the world.

> Through naming comes knowing; we grasp an object, mentally, by giving it a name—hension, prehension, apprehension. And thus through language create a whole world, corresponding to the other world out there. Or we trust that it corresponds. Or perhaps, like a German poet, we cease to care, becoming more concerned with the naming than with the things named; the former becomes more real than the latter. And so in the end the world is lost again. No, the world remains—those unique, particular, incorrigibly individual junipers and sandstone monoliths—and it is we who are lost. Again. Round and round, through the endless labyrinth of thought—the maze. (Edward Abbey [1, pp. 288–289].)

4. Comments on the logical approach

The basic idea underlying the logical approach to AI is simple, but attempts to use it have resulted in several additional important insights.

4.1. The importance of conceptualization

The most important part of "the AI problem" involves inventing an appropriate conceptualization (intended model). It is not easy for a designer to squeeze his intuitive and commonsense ideas about the world into a coherent conceptualization involving objects, functions, and relations. Although this exercise has been carried out for several limited problem domains (most notably those to which expert systems have been successfully applied), there are some particularly difficult subjects to conceptualize. Among these are liquids and other "mass substances," processes, events, actions, beliefs, time, goals, intentions, and plans. Some researchers feel that the *frame problem*, for example, arises as it does as an artifact of an inappropriate (state-based) conceptualization of change [17]. Others feel that change must involve the notion of time (instead of the notion of state) [52]. Conceptualizing the "cognitive state" of intelligent agents has been the subject of recent intense study. (See, for example, [8] for a treatment of the intentions of agents and [24, 40] for treatments of the knowledge and beliefs of agents.) Interestingly, many of the most difficult conceptualization problems arise when attempting to express knowledge about the everyday, "commonsense" world (see [20, 21]). AI researchers join company with philosophers who have also been attempting to formalize some of these ideas.

Choosing to use first-order predicate calculus as a representation language does not relieve us of the chore of deciding *what* to say in that language. Deciding what to say is harder than designing the language in which to say it! The logical approach to AI carries with it no special insights into what conceptualizations to use. (Logic is often criticized for providing *form* but not *content*. Of course!)

It is important to stress that these conceptualization problems do not arise simply as an undesirable side effect of the use of logic. They must be confronted and resolved by any approach that attempts to represent knowledge of the world by sentence-like, declarative structures. The fact that these problems are exposed quite clearly in the coherent framework provided by the logical approach should be counted as an advantage.

4.2. *Sound and unsound inferences*

Another important observation concerns the subject of sound inference. Logicists are sometimes criticized for their alleged dependence on deduction. Much human thought, the critics rightly claim, involves leaps of intuition, inductive inference, and other guessing strategies that lie outside the realm of sound inference. There are two things that can be said about such criticism.

First, logicists regard sound inference as an important, but not the only, component of reasoning. We must be careful to note the circumstances under which both sound and unsound inferences might appropriately be used. Recall that the set of sentences Δ (with which a designer endows a machine) implicitly defines a set of models. Either the designer actually has some subset of these models in mind (as his guess about what the world is) or he is completely unbiased about which of the models might represent the world. If he really is unbiased, nothing other than sound inference would be desired by the designer. Any deduced sentence ϕ had better be logically entailed by Δ; if there are some models of Δ, for example, that are not models of ϕ, and if the designer wanted the machine to conclude ϕ, then he wouldn't have been completely unbiased about which of the models of Δ represented the world.

If the designer has some subset of the models of Δ in mind, and if (for one reason or another) he could not specify this subset by enlarging Δ, then there are circumstances under which unsound inference might be appropriate. For example, the designer may have some preference order over the models of Δ. He may want to focus, for example, on the minimal models (according to the preference order). These minimal models may be better guesses, in the designer's mind, about the real world than would be the other models of Δ. In justification. For example, circumscription is motivated by minimal-model entailment and thus might be called a "principled" inference even though not a sound one.

4.3. *Efficiency and semantic attachment to partial models*

Earlier, we mentioned that it was fortunate that sound inference techniques existed because it is impossible in most situations to check that all the models of Δ were also models of some formula ϕ. This "good fortune" is somewhat illusory however, because finding deductions is in general intractable and for many practical applications unworkably inefficient. Some people think that the inefficiency of the logical approach disqualifies it from serious consideration as a design strategy for intelligent machines.

There are several things to be said about logic and efficiency. First, it seems incontestable that knowledge can be brought to bear on a problem more efficiently when its use is tailored to the special features of that problem. When knowledge is encoded in a fashion that permits many different uses, several possible ways in which to use it may have to be tried in any given situation, and the resulting search process takes time. A price does have to be paid for generality, and the logical approach, it seems, pays a runtime cost to save accumulated design costs.

But even so, much progress has been made in making inference processes more efficient and practical for large problems. Stickel has developed one of the most powerful first-order-logic theorem provers [56, 57]. Several resolution refutation systems have been written that are able to solve large, nontrivial reasoning problems, including some open problems in mathematics [59, 61]. Many large-scale AI systems depend heavily on predicate calculus representations and reasoning methods. Among the more substantial of these are TEAM, a natural language interface to databases [14]; DART, a program for equipment design and repair [11]; and KAMP, a program that generates English sentences [3].

A very important technique for achieving efficiency in the context of the logical approach involves augmenting theorem-proving methods with calculations on model-like structures. Often, calculations on models are much more efficient than are inference processes, and we would be well advised to include them as part of a machine's reasoning apparatus.

We mentioned that seldom does a designer make explicit his guess about the world, the intended model. The set of models is implicitly defined by the set of sentences in Δ. Sometimes, however, it is possible to be explicit about at least part of the intended model. That is, we might be able to construct a part of the model as list structure and programs in, say, LISP. For example, we can represent objects as LISP atoms, functions as LISP functions, and relations as LISP predicates. In such cases we can perform reasoning by computations with explicitly as part of the language [5]). Various LISP ordering predicates combined with appropriate directed-graph data structures are useful for representing transitive binary relations.

4.4. *Reification of theories*

Sometimes we will want our machines to reason about (rather than with) the sentences in its knowledge base. We may, for example, want them to reason about the lengths of sentences or about their complexity. Our conceptualizations will thus have to acknowledge that things called sentences exist in the world. Conferring *existence* on abstract concepts (such as sentences) is often called *reification*.

We might reify whole theories. This will allow us to say, for example, that some Δ_1 is more appropriate than is some Δ_2 when confronted with problems of diagnosing bacterial infections. Scientists are used to having different—even contradictory—theories to explain reality: quantum physics, Newtonian mechanics, relativity, wave theories of light, particle theories of light, and so on. Each is useful in certain circumstances. Although scientists search for a uniform, all-embracing, and consistent picture of reality, historically they have had to settle for a collection of somewhat different theories. There is nothing in the logicist approach that forces us, as machine designers, to use just *one* conceptualization of the world. There is no reason to think AI would be any more successful at that goal than scientists have been!

When theories are reified, *metatheory* (that is, a theory about theories) can be used to make decisions about which local theory should be used in which circumstances. For example, the metatheory might contain a predicate calculus statement having an intended meaning something like: "When planning a highway route, use the theory that treats roads as edges in a graph (rather than, for example, as solid objects made of asphalt or concrete)". Metatheory can also provide information to guide the inference procedures operating over local theories. For example, we might want to say that when two inferences are possible in some Δ_1, the inference that results in the most general conclusion should be preferred. Using metatheory to express knowledge about how to control inference is consistent with the logicists' desire to put as much knowledge as possible in declarative form (as opposed to "building it in" to the functions *mem* and *act*).

Weyhrauch [58] has pointed out that the process of semantic attachment in a metatheory can be particularly powerful. Commonly, even when no semantic attachments are possible to speed reasoning in a theory, the problem at hand can be dispatched efficiently by appropriate semantic attachment in the metatheory.

Some critics of the logical approach have claimed that since *anything* can be said in the metatheory, its use would seem to be a retreat to the same ad hoc tricks used by less disciplined AI researchers. But we think there are generally useful things to say in the metatheory that are not themselves problem dependent. That is, we think that knowledge about how to use knowledge can itself be expressed as context-free, declarative sentences. (Lenat's work has uncovered the best examples of generally useful statements about how to use knowledge [25–27].)

4.5. *Other observations*

Even though they frequently call the sentences in their knowledge bases *axioms*, logicists are not necessarily committed to represent knowledge by a minimal set of sentences. Indeed, some (or even most) of the sentences in Δ may be derivable from others. Since the "intelligence" of an agent depends on how much usable declarative knowledge it has, we agree completely with those who say "In the knowledge lies the power." We do not advocate systems that rely on search-based derivations of knowledge when it is possible to include the needed knowledge explicitly in the knowledge base. The use of very large knowledge bases, of course, presupposes efficient retrieval and indexing techniques.

The occasional criticism that logicists depend too heavily on their inference methods and not on the knowledge base must simply result from a misunderstanding of the goals of the logical approach. As has already been pointed out, logicists strive to make the inference process as uniform and domain independent as possible and to represent all knowledge (even the knowledge about how to use knowledge) declaratively.

5. Challenging problems

5.1. *Language and the world*

Few would deny that intelligent machines must have some kind of characterization or model of the world they inhabit. We have stressed that the main feature of machines designed using the logical approach is that they describe their worlds by *language*. Is language (any language) adequate to the task? As the writer Edward Abbey observed [1, p. x]:

> Language makes a mighty loose net with which to go fishing for simple facts, when facts are infinite.

A designer's intuitive ideas about the world are often difficult to capture in a conceptualization that can be described by a finite set of sentences. Usually these intuitive ideas are never complete at the time of design anyway, and the conceptualization expands making it difficult for the sentences to catch up.

John McCarthy humorously illustrates this difficulty by imagining how one might formulate a sentence that says that under certain conditions a car will start. In English we might say, for example: "If the fuel tank is not empty and if you turn the ignition key, the car will start." But this simple sentence is not true of a world in which the carburetor is broken, or in which the fuel tank (while not empty) is full of water, or in which the exhaust pipe has a potato stuck in it, or . . . . Indeed, it seems there might be an infinite number of *qualifications* that would need to be stated in order to make such a sentence true (in the world the designer has in mind—or comes to have in mind). Of course, just what it means for a designer to have a world in mind is problematical; he probably didn't even think of the possibility of the potato in the tailpipe until it was mentioned by someone else who happened to conceive of such a world.

There seem to be two related problems here. One is that we would like to have and use approximate, simple conceptualizations even when our view of the world would permit more accurate and detailed ones. The approximate ones are often sufficient for our purposes. Thus, even though we know full well that the carburetor must be working in order for a car to start, in many situations for which we want to reason about the car starting we don't need to know about the carburetor and can thus leave it out of our conceptualization. Using theories (Δ's) corresponding to approximate conceptualizations and successive refinements of them would seem to require the ability to have several such at hand and a metatheory to decide when to use which.

Another problem is that even the most detailed and accurate conceptualization may need to be revised as new information becomes available. Theories must be revisable to accomodate the designer's changing view of the world. As the machine interacts with its world, it too will learn new information which will in some cases add to its theory and in other cases require it to be modified.

Science has similar problems. Scientists and engineers knowingly and usefully employ approximate theories—such as frictionless models. Furthermore, all of our theories of the physical world are falsifiable, and, indeed, we expect scientific progress to falsify the theories we have and to replace them by others. When we conclude something based on a current physical theory, we admit the dependence of the conclusion on the theory and modify the theory if the conclusion is contradicted by subsequent facts. Those who would argue that logical languages are inappropriate for representing synthetic or contingent knowledge about the world [39] would also seem to have to doubt the utility of any of the languages that science uses to describe and predict reality. Merely because our conceptualization of the world at any stage of our progress toward understanding it may (inevitably will!) prove to be inaccurate does not mean that this conceptualization is not in the meantime useful.

Some AI researchers have suggested techniques for making useful inferences from an approximate, but not inaccurate, theory. We say that a theory is not inaccurate if its models include the world as conceived by the designer. If a theory is to be not inaccurate, it is typically impossible or overly cumbersome to include the universal statements needed to derive useful sound conclusions.

We illustrate the difficulty by an example. Suppose that we want our machine to decide whether or not an apple is edible. If Δ is to be not inaccurate, we cannot include in it the statement $(\forall x)\mathrm{Apple}(x) \wedge \mathrm{Ripe}(x) \supset \mathrm{Edible}(x)$ in the face of known exceptions such as Wormy(x) or Rotten(x). (We trust that the reader understands that the mnemonics we use in our examples must be backed up by sufficient additional statements in Δ to insure that these mnemonics are constrained to have roughly their intended meanings.) Suppose we cannot conclude from Δ that a given apple, say the apple denoted by apple1 is wormy or rotten; then we may want to conclude (even non-soundly) Edible(Apple1). If later, it is learned (say through sensory inputs) that Rotten(Apple1), then we must withdraw the earlier conclusion Edible(Apple1). The original inference is called *defeasible* because it can be defeated by additional information. Making such inferences involves what is usually called *nonmonotonic reasoning*. (Ordinary logical reasoning is monotonic in the sense that the set of conclusions that can be drawn from a set of sentences is not diminished if new sentences are added.)

Several researchers have proposed frameworks and techniques for nonmonotonic reasoning. McDermott and Doyle [37, 38] have developed a *nonmonotonic logic*. Reiter [46] has proposed inference rules (called *default rules*) whose applicability to a set of sentences Δ depends on what is *not* in Δ as well as what is. McCarthy [34] advocates the use of *circumscription* based on minimal models. Ginsberg [13] uses multiple (more than two) truth values to represent various degrees of knowledge. We will briefly describe one of these approaches, that based on minimal models, in order to illustrate what can be done. (See [47] for a thorough survey.)

Consider the general rule $(\forall x)Q(x) \supset P(x)$. We may know that this rule is not strictly correct without additional qualifications, and thus it cannot be included in a machine's knowledge base without making the knowledge base inaccurate. But we may want to use something like this rule to express the fact that "typically" all objects satisfying property Q also satisfy property P. Or we may want to use the rule in a system that can tolerate qualifications to be added later.

One way to hedge the rule (to avoid inaccuracy) is to introduce the concept of *abnormality*, denoted by the relation constant Ab [35]. Then we can say that all objects that are not abnormal and that satisfy property Q also satisfy property P:

$$(\forall x)Q(x) \wedge \neg \mathrm{Ab}(x) \supset P(x) .$$

Which objects are abnormal and which are not (if we know these facts) can be specified by other sentences in Δ. For example we may know that the objects denoted by A and B are abnormal: $\mathrm{Ab}(A) \wedge \mathrm{Ab}(B)$.

The frame problem has been thoroughly treated in the literature. (See [7] and [44] for collections of articles. The latter collection includes several that discuss the problem from the standpoints of philosophy and cognitive psychology.) In attempting to deal with the frame problem in their system called STRIPS, Fikes and Nilsson [10] described the effects of a machine's actions by listing those relations that were changed by the action. They assumed that those relations not mentioned were not changed. Hayes [17, 18] introduced the notion of *histories* in an attempt to define a conceptualization in which the frame problem was less severe. McCarthy [35] and Reiter [46] proposed nonmonotonic reasoning methods for dealing with the frame problem. In the language of circumscription, their approaches assumed minimal changes consistent with the relations that were known to change. However, Hanks and McDermott [15] showed that a straightforward application of circumscription does not produce results strong enough to solve the frame problem. In response, Lifschitz [30] introduced a variant called *pointwise circumscription*. He also proposed reconceptualization of actions and their effects that permits the use of ordinary circumscription in solving the frame problem and the qualification problem [31]. Shoham [51] proposed an alternative minimization method related to circumscription, called *chronological ignorance*.

Although the frame problem has been extensively studied, it remains a formidable conceptual obstacle to the development of systems that must act in a changing world. This obstacle is faced by all such systems—even those whose knowledge about the world is represented in procedures. The designer of any intelligent machine must make assumptions (at least implicit ones) about how the world changes in response to the actions of the machine if the machine is to function effectively.

5.3. *Uncertain knowledge*

When one is uncertain about the world, one cannot specify precisely which relations hold in the world. Nevertheless, one might be able to say that at least one of a set of relations holds. Logical disjunctions permit us to express that kind of uncertain knowledge.

Logical representations (with their binary truth values) would seem to be inadequate for representing other types of uncertain knowledge. How do we say, for example, "It is likely that it will be sunny in Pasadena on New Year's day"? We could, of course embed probability information itself in the sentence, and this approach and others have been followed. Attempts to fuzz the crisp true/false semantics of logical languages have led to an active AI research subspecialty [23, 43, 53].

The approach followed by [41], for example, is to imagine that a probability value is associated with each of a set of possible conceptualizations (interpretations). The machine designer makes this assignment implicitly by composing a

9 780262 610759
90000

Cover art: *Algen2*, 1990.

Reprinted from a special issue of
Artificial Intelligence: An International Journal

FOUNDATIONS OF ARTIFICIAL INTELLIGENCE

edited by David Kirsh

Have the classical methods and ideas of AI outlived their usefulness? *Foundations of Artificial Intelligence* critically evaluates the fundamental assumptions underpinning the dominant approaches to AI. In these eleven contributions, theorists historically associated with each position identify the basic tenets of their position. They discuss the underlying principles, describe the natural types of problems and tasks in which their approach succeeds, explain where its power comes from, and what its scope and limits are. Theorists generally skeptical of these positions evaluate the effectiveness of the method or approach and explain why it works—to the extent they believe it does—and why it eventually fails.

Among the key questions discussed are, What is a theory in AI? What are AI's central problems? Does intelligence require declarative knowledge and some form of reasoning-like computation? Or can AI be achieved by complex control systems operating without robust concepts? Can central cognition be studied and tested without prior theories of perception and motor control? Or do the two types of theories—central or peripheral—interpenetrate? Can the trajectory of information states created during cognition be described in English or some logico-mathematical version of English? Or will we need a subconceptual level? Is there a single architecture underlying all cognition? Or is intelligence the product of thousands of specialized subsystems?

David Kirsh is Assistant Professor in the Department of Cognitive Science at the University of California, San Diego.

A Bradford Book

The MIT Press
Massachusetts Institute of Technology
Cambridge, Massachusetts 02142

KIRFP
0-262-61075-2