What is a Knowledge Representation?[1]

Randall Davis
MIT AI Lab

Howard Shrobe
MIT AI Lab and Symbolics, Inc.

Peter Szolovits
MIT Lab for Computer Science

Abstract

Although knowledge representation is one of the central and in some ways most familiar concepts in AI, the most fundamental question about it—What is it?—has rarely been answered directly. Numerous papers have lobbied for one or another variety of representation, other papers have argued for various properties a representation should have, while still others have focused on properties that are important to the notion of representation in general.

In this paper we go back to basics to address the question directly. We believe that the answer can best be understood in terms of five important and distinctly different roles that a representation plays, each of which places different and at times conflicting demands on the properties a representation should have.

We argue that keeping in mind all five of these roles provides a usefully broad perspective that sheds light on some longstanding disputes and can invigorate both research and practice in the field.

[1] This report describes research done at the Artificial Intelligence Laboratory and the Laboratory for Computer Science at MIT. Support for this work was received from the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124, the National Library of Medicine under grant R01 LM 04493, the National Heart, Lung and Blood Institute under grant R01 HL 33041, Digital Equipment Corporation and the General Dynamics Corp.

This paper was published as R. Davis, H. Shrobe, and P. Szolovits. What is a Knowledge Representation? *AI Magazine*, 14(1):17-33, 1993.

Contents

1 Introduction **1**

2 Terminology and Perspective **2**

3 What is a Knowledge Representation? **2**

3.1 Role I: A KR is a Surrogate . . . 2

3.1.1 Imperfect surrogates mean incorrect inferences are inevitable . 3

3.2 Role II: A KR is a Set of Ontological Commitments . . . 4

3.2.1 Commitment begins with the earliest choices . . . 5

3.2.2 The commitments accumulate in layers . . . 6

3.2.3 Reminder: A KR is not a data structure . . . 6

3.3 Role III: A KR is a Fragmentary Theory Of Intelligent Reasoning . . 7

3.3.1 What is intelligent reasoning? . . . 8

3.3.1.1 Intelligent reasoning: the logical view and the psychological view . . . 9

3.3.2 Which inferences are sanctioned? . . . 11

3.3.2.1 Form and content of the answers . . . 12

3.3.3 Which inferences are recommended? . . . 13

3.4 Role IV: A KR is a Medium for Efficient Computation . . . 15

3.5 Role V: A KR is a Medium of Human Expression . . . 16

4 Consequences For Research and Practice **17**

4.1 Consequence for Practice: Characterizing the Spirit of a Representation 17

4.1.1 The spirit should be *indulged*, not *overcome* . . . 18

4.2 Consequence for Research: Representation and Reasoning are Intertwined . . . 19

4.3 Consequence for Research: Combining Representations . . . 20

4.4 Consequence for Research: Arguments About Formal Equivalence . . 21

4.5 Consequence for Research: All Five Roles Matter . . . 22

5 The Goal of KR Research **23**

6 Summary **23**

1 Introduction

What is a knowledge representation? We argue that the notion can best be understood in terms of five distinct roles it plays, each crucial to the task at hand:

- A knowledge representation (KR) is most fundamentally a **surrogate**, a substitute for the thing itself, used to enable an entity to determine consequences by thinking rather than acting, i.e., by reasoning about the world rather than taking action in it.

- It is a **set of ontological commitments**, i.e., an answer to the question: In what terms should I think about the world?

- It is a **fragmentary theory of intelligent reasoning**, expressed in terms of three components: (i) the representation's fundamental conception of intelligent reasoning; (ii) the set of inferences the representation **sanctions**; and (iii) the set of inferences it **recommends**.

- It is a **medium for pragmatically efficient computation**, i.e., the computational environment in which thinking is accomplished. One contribution to this pragmatic efficiency is supplied by the guidance a representation provides for organizing information so as to facilitate making the recommended inferences.

- It is a **medium of human expression**, i.e., a language in which we say things about the world.

Understanding the roles and acknowledging their diversity has several useful consequences. First, each role requires something slightly different from a representation; each accordingly leads to an interesting and different set of properties we want a representation to have.

Second, we believe the roles provide a framework useful for characterizing a wide variety of representations. We suggest that the fundamental "mindset" of a representation can be captured by understanding how it views each of the roles, and that doing so reveals essential similarities and differences.

Third, we believe that some previous disagreements about representation are usefully disentangled when all five roles are given appropriate consideration. We demonstrate this by revisiting and dissecting the early arguments concerning frames and logic.

Finally, we believe that viewing representations in this way has consequences for both research and practice. For research, this view provides one direct answer to a question of fundamental significance in the field. It also suggests adopting a broad perspective on what's important about a representation, and it makes the case that one significant part of the representation endeavor—capturing and representing the richness of the natural world—is receiving insufficient attention. We believe this view can also improve practice by reminding practitioners about the inspirations that are the important sources of power for a variety of representations.

2 Terminology and Perspective

Two points of terminology will assist in our presentation. First, we use the term *inference* in a generic sense, to mean any way to get new expressions from old. We are only rarely talking about sound logical inference and when doing so refer to that explicitly.

Second, to give them a single collective name, we refer to the familiar set of basic representation tools like logic, rules, frames, semantic nets, etc., as *knowledge representation technologies*.

It will also prove useful to take explicit note of the common practice of building knowledge representations in multiple levels of languages, typically with one of the knowledge representation technologies at the bottom level. Hayes' ontology of liquids [12], for example, is at one level a representation composed of concepts like `pieces` of space, that have `portals`, `faces`, `sides`, etc. The language at the next more primitive (and as it turns out, bottom) level is first order logic, where, for example, $\mathtt{In}(\mathtt{s}_1,\mathtt{s}_2)$ is a relation expressing that space $\mathtt{s}_1$ is contained in $\mathtt{s}_2$.

This view is useful in part because it allows our analysis and discussion to concentrate largely on the KR technologies. As the primitive representational level at the foundation of KR languages, they encounter all of the issues central to knowledge representation of any variety. They are also useful exemplars because they are widely familiar to the field and there is a substantial body of experience with them to draw on.

3 What is a Knowledge Representation?

Perhaps the most fundamental question about the concept of knowledge representation is, What *is* it? We believe that the answer is best understood in terms of the five fundamental roles that it plays.

3.1 Role I: A KR is a Surrogate

Any intelligent entity that wishes to reason about its world encounters an important, inescapable fact: reasoning is a process that goes on internally, while most things it wishes to reason about exist only externally. A program (or person) engaged in planning the assembly of a bicycle, for instance, may have to reason about entities like wheels, chains, sprockets, handle bars, etc., yet such things exist only in the external world.

This unavoidable dichotomy is a fundamental rationale and role for a representation: it functions as a surrogate inside the reasoner, a stand-in for the things that exist in the world. Operations on and with representations substitute for operations on the real thing, i.e., substitute for direct interaction with the world. In this view reasoning itself is in part a surrogate for action in the world, when we can not or do not (yet) want to take that action.[2]

[2]Conversely, action can substitute for reasoning. This dualism offers one way to understand

Viewing representations as surrogates leads naturally to two important questions. The first question about any surrogate is its intended identity: what is it a surrogate for? There must be some form of correspondence specified between the surrogate and its intended referent in the world; the correspondence is the semantics for the representation.

The second question is fidelity: how close is the surrogate to the real thing? What attributes of the original does it capture and make explicit, and which does it omit? Perfect fidelity is in general impossible, both in practice and in principle. It is impossible in principle because any thing other than the thing itself is necessarily different from the thing itself (in location if nothing else). Put the other way around, the only *completely accurate* representation of an object is the object itself. All other representations are inaccurate; they inevitably contain simplifying assumptions and possibly artifacts.

Two minor elaborations extend this view of representations as surrogates. First, it appears to serve equally well for intangible objects as it does for tangible objects like gear wheels: Representations function as surrogates for abstract notions like actions, processes, beliefs, causality, categories, etc., allowing them to be described inside an entity so it can reason about them. Second, formal objects can of course exist inside the machine with perfect fidelity: Mathematical entities, for example, can be captured exactly, precisely because they are formal objects. Since almost any reasoning task will encounter the need to deal with natural objects (i.e., those encountered in the real world) as well as formal objects, imperfect surrogates are pragmatically inevitable.

3.1.1 Imperfect surrogates mean incorrect inferences are inevitable

Two important consequences follow from the inevitability of imperfect surrogates. One consequence is that in describing the natural world, we must inevitably lie, by omission at least. At a minimum we must omit some of the effectively limitless complexity of the natural world; our descriptions may in addition introduce artifacts not present in the world.

The second and more important consequence is that all sufficiently broad-based reasoning about the natural world *must* eventually reach conclusions that are incorrect, independent of the reasoning process used and independent of the representation employed. Sound reasoning cannot save us: If the world model is somehow wrong (and it must be) some conclusions will be incorrect, no matter how carefully drawn. A better representation cannot save us: all representations are imperfect and any imperfection can be a source of error.

The significance of the error may of course vary; indeed much of the art of selecting a good representation is in finding one that minimizes (or perhaps even eliminates) error for the specific task at hand. But the unavoidable imperfection of surrogates

the relation between traditional symbolic representations and the situated action approach, which argues that action can be linked directly to perception, without the need for intermediating symbolic representations.

means we can supply at least one guarantee for any entity reasoning in any fashion about the natural world: if it reasons long enough and broadly enough, it is *guaranteed* to err.

Drawing only sound inferences thus does not free reasoning from error; it can only ensure that inference is not the source of that error. Given that broad based reasoning will inevitably be wrong, the step from sound inference to other models of inference is thus not a move from total accuracy to error, but is instead a question of balancing off the possibility of one more source of error against the gains (e.g., efficiency) it may offer.

We do not suggest that unsound reasoning ought to be embraced casually, but do claim that, given the inevitability of error even with sound reasoning, it makes sense to evaluate pragmatically the relative costs and benefits that come from using both sound and unsound reasoning methods.

3.2 Role II: A KR is a Set of Ontological Commitments

If, as we have argued, all representations are imperfect approximations to reality, each approximation attending to some things and ignoring others, then in selecting any representation we are in the very same act unavoidably making a set of decisions about how and what to see in the world. That is, selecting a representation means making a set of ontological commitments.[3] The commitments are in effect a strong pair of glasses that determine what we can see, bringing some part of the world into sharp focus, at the expense of blurring other parts.

These commitments and their focusing/blurring effect are not an incidental side effect of a representation choice; they are of the essence: a KR *is* a set of ontological commitments. It is *unavoidably* so because of the inevitable imperfections of representations. It is *usefully* so because judicious selection of commitments provides the opportunity to focus attention on aspects of the world we believe to be relevant.

The focusing effect is an essential part of what a representation offers, because the complexity of the natural world is overwhelming. We (and our reasoning machines) need guidance in deciding what in the world to attend to and what to ignore. The glasses supplied by a representation can provide that guidance: In telling us what and how to see, they allow us to cope with what would otherwise be untenable complexity and detail. Hence the ontological commitment made by a representation can be one of the most important contributions it offers.

There is a long history of work attempting to build good ontologies for a variety of task domains, including early work on an ontology for liquids [12], the lumped

[3] The phrase "ontological commitment" is perhaps not precisely correct for what we have in mind here, but it is the closest available approximation. While ontology is, strictly speaking, concerned with what *exists* in the world, we have phrased this section carefully in terms of how to *view* the world, purposely sidestepping many standard thorny philosophical issues surrounding claims of what exists. A second way around the issue is to note that the "world" we are interested in capturing is the world inside the mind of some intelligent human observer (e.g., a physician, engineer, etc.), in which case it can plausibly be argued that in *that* world rules, prototypes, etc., do exist.

element model widely used in representing electronic circuits (e.g., [3]), as well as ontologies for time, belief, and even programming itself. Each of these offers a way to see some part of the world.

The lumped element model, for instance, suggests that we think of circuits in terms of `components` with `connections` between them, with `signals` flowing instantaneously along the `connections`. This is a useful view, but not the only possible one. A different ontology arises if we need to attend to the electrodynamics in the device: Here signals propagate at finite speed and an object (like a resistor) that was previously viewed as a single component with an I/O behavior may now have to be thought of as an extended medium through which an electromagnetic wave flows.

Ontologies can of course be written down in a wide variety of languages and notations (e.g., logic, LISP, etc.); the essential information is not the form of that language but the content, i.e., the set of concepts offered as a way of thinking about the world. Simply put, the important part is notions like `connections` and `components`, not whether we choose to write them as predicates or LISP constructs.

The commitment we make by selecting one or another ontology can produce a sharply different view of the task at hand. Consider the difference that arises in selecting the lumped element view of a circuit rather than the electrodynamic view of the same device. As a second example, medical diagnosis viewed in terms of rules (e.g., MYCIN) looks substantially different from the same task viewed in terms of frames (e.g., INTERNIST). Where MYCIN sees the medical world as made up of empirical associations connecting symptom to disease, INTERNIST sees a set of prototypes, in particular prototypical diseases, to be matched against the case at hand.

3.2.1 Commitment begins with the earliest choices

The INTERNIST example also demonstrates that there is significant and unavoidable ontological commitment even at the level of the familiar representation technologies. Logic, rules, frames, etc., each embody a viewpoint on the kinds of things that are important in the world. Logic, for instance, involves a (fairly minimal) commitment to viewing the world in terms of individual entities and relations between them. Rule-based systems view the world in terms of attribute-object-value triples and the rules of plausible inference that connect them, while frames have us thinking in terms of prototypical objects.

Each of these thus supplies its own view of what is important to attend to, and each suggests, conversely, that anything not easily seen in those terms may be ignored. This is of course not guaranteed to be correct, since anything ignored may later prove to be relevant. But the task is hopeless in principle—every representation ignores something about the world—hence the best we can do is start with a good guess. The existing representation technologies supply one set of guesses about what to attend to and what to ignore. Selecting any of them thus involves a degree of ontological commitment: the selection will have a significant impact on our perception of and approach to the task, and on our perception of the world being modeled.

3.2.2 The commitments accumulate in layers

The ontologic commitment of a representation thus begins at the level of the representation technologies and accumulates from there. Additional layers of commitment are made as we put the technology to work. The use of frame-like structures in INTERNIST offers an illustrative example. At the most fundamental level, the decision to view diagnosis in terms of frames suggests thinking in terms of prototypes, defaults, and a taxonomic hierarchy. But prototypes of what, and how shall the taxonomy be organized?

An early description of the system [21] shows how these questions were answered in the task at hand, supplying the second layer of commitment:

> The knowledge base underlying the INTERNIST system is composed of two basic types of elements: disease entities and manifestations.... [It] also contains a ... hierarchy of disease categories, organized primarily around the concept of organ systems, having at the top level such categories as "liver disease," "kidney disease," etc.

The prototypes are thus intended to capture prototypical diseases (e.g., a "classic case" of a disease), and they will be organized in a taxonomy indexed around organ systems. This is a sensible and intuitive set of choices but clearly not the only way to apply frames to the task; hence it is another layer of ontological commitment.

At the third (and in this case final) layer, this set of choices is instantiated: which diseases will be included and in which branches of the hierarchy will they appear? Ontologic questions that arise even at this level can be quite fundamental. Consider for example determining which of the following are to be considered *diseases* (i.e., abnormal states requiring cure): alcoholism, homosexuality, and chronic fatigue syndrome. The ontologic commitment here is sufficiently obvious and sufficiently important that it is often a subject of debate in the field itself, quite independent of building automated reasoners.

Similar sorts of decisions have to be made with all the representation technologies, because each of them supplies only a first order guess about how to see the world: they offer a way of seeing but don't indicate how to instantiate that view. As frames suggest prototypes and taxonomies but do not tell us which things to select as prototypes, rules suggest thinking in terms of plausible inferences, but don't tell us which plausible inferences to attend to. Similarly logic tells us to view the world in terms of individuals and relations, but does not specify which individuals and relations to use.

Commitment to a particular view of the world thus starts with the choice of a representation technology, and accumulates as subsequent choices are made about how to see the world in those terms.

3.2.3 Reminder: A KR is not a data structure

Note that at each layer, even the first (e.g., selecting rules or frames), the choices being made are about representation, not data structures. Part of what makes a language representational is that it carries meaning [13], i.e., there is a correspondence between its constructs and things in the external world. That correspondence in turn carries with it constraint.

A semantic net, for example, is a representation, while a graph is a data structure. They are different kinds of entities, even though one is invariably used to implement the other, precisely because the net has (should have) a semantics. That semantics will be manifest in part because it constrains the network topology: a network purporting to describe family memberships as we know them cannot have a cycle in its `parent` links, while graphs (i.e., data structures) are of course under no such constraint and may have arbitrary cycles.

While every representation must be *implemented* in the machine by some data structure, the representational property is in the correspondence to something in the world and in the constraint that correspondence imposes.

3.3 Role III: A KR is a Fragmentary Theory Of Intelligent Reasoning

The third role for a representation is as a fragmentary theory of intelligent reasoning. This role comes about because the initial conception of a representation is typically motivated by some insight indicating how people reason intelligently, or by some belief about what it means to reason intelligently at all.

The theory is fragmentary in two distinct senses: (*i*) the representation typically incorporates only part of the insight or belief that motivated it, and (*ii*) that insight or belief is in turn only a part of the complex and multi-faceted phenomenon of intelligent reasoning.

A representation's theory of intelligent reasoning is often implicit, but can be made more evident by examining its three components: (*i*) the representation's fundamental conception of intelligent inference; (*ii*) the set of inferences the representation sanctions; and (*iii*) the set of inferences it recommends.

Where the sanctioned inferences indicate what can be inferred at all, the recommended inferences are concerned with what should be inferred. (Guidance is needed because the set of sanctioned inferences is typically far too large to be used indiscriminantly.) Where the ontology we examined earlier tells us how to see, the recommended inferences suggest how to reason.

These components can also be seen as the representation's answers to three corresponding fundamental questions: (*i*) What does it mean to reason intelligently? (*ii*) What can we infer from what we know? and (*iii*) What ought we to infer from what we know? Answers to these questions are at the heart of a representation's spirit and mindset; knowing its position on these issues tells us a great deal about it.

We begin with the first of these components, examining two of several fundamentally different conceptions of intelligent reasoning that have been explored in AI. Those conceptions and their underlying assumptions demonstrate the broad range of views on the question and set important context for the remaining components.

3.3.1 What is intelligent reasoning?

What are the essential, defining properties of intelligent reasoning?

As a consequence of the relative youth of AI as a discipline, insights about the nature of intelligent reasoning have often come from work in other fields. Five fields—mathematical logic, psychology, biology, statistics, and economics—have provided the inspiration for five distinguishable notions of what constitutes intelligent reasoning (Table I).

One view, historically derived from mathematical logic, makes the assumption that intelligent reasoning is some variety of formal calculation, typically *deduction*; the modern exemplars of this view in AI are the logicists. A second view, rooted in work in psychology, sees reasoning as a characteristic *human behavior* and has given rise to both the extensive work on human problem solving and the large collection of knowledge-based systems.

A third approach, loosely rooted in biology, takes the view that the key to reasoning is the architecture of the machinery that accomplishes it, hence reasoning is a characteristic stimulus/response behavior that emerges from *parallel interconnection of a large collection of very simple processors.* Researchers working on several varieties of connectionism are the current descendants of this line of work. A fourth approach, derived from probability theory, adds to logic the notion of uncertainty, yielding a view in which reasoning intelligently means *obeying the axioms of probability theory.* A fifth view, from economics, adds the further ingredient of values and preferences, leading to a view of intelligent reasoning defined by adherence to the tenets of *utility theory.*

MATHEMATICAL LOGIC	PSYCHOLOGY	BIOLOGY	STATISTICS	ECONOMICS
Aristotle				
Descartes				
Boole	James		Laplace	Bentham Pareto
Frege Peano			Bernoulli	Friedman
	Hebb	Lashley	Bayes	
Goedel	Bruner	Rosenblatt		
Post	Miller	Ashby	Tversky,	Von Neumann
Church	Newell,	Lettvin	Kahneman	Simon
Turing	Simon	McCulloch, Pitts		Raiffa
Davis		Heubel, Weisel		
Putnam				
Robinson				
LOGIC PROLOG	SOAR KBS, FRAMES	CONNECTIONISM	CAUSAL NETWORKS	RATIONAL AGENTS

Table I. Views of intelligent reasoning and their intellectual origins

Briefly exploring the historical development of the first two of these views will illustrate the different conceptions they have of the fundamental nature of intelligent reasoning and will demonstrate the deep-seated differences in mindset that arise as a consequence.

3.3.1.1 Intelligent reasoning: the logical view and the psychological view

Consider first the tradition that surrounds mathematical logic as a view of intelligent reasoning. This view has its historical origins in Aristotle's efforts to accumulate and catalog the syllogisms, in an attempt to determine what should be taken as a convincing argument.[4] The line continues with Descartes, whose analytic geometry showed that Euclid's work, apparently concerned with the stuff of pure thought (lines of zero width, perfect circles of the sorts only the gods could make), could in fact be married to algebra, a form of calculation, something mere mortals could do.

[4]Note that even at the outset there is a hint of the idea that the desired form of reasoning might be describable in a set of formal rules.

By the time of Leibnitz the agenda is quite specific and telling: he sought nothing less than a "calculus of thought," one that would permit the resolution of all human disagreement with the simple invocation "Let us compute." By this time there is a clear and concrete belief that, as Euclid's once god-like and unreachable geometry could be captured with algebra, so some (or perhaps any) variety of that ephemeral stuff called thought might be captured in *calculation*, specifically logical deduction.

In the 19th century Boole provided the basis for propositional calculus in his "Laws of Thought;" later work by Frege and Peano provided additional foundation for the modern form of predicate calculus. Work by Davis, Putnam and Robinson in the 20th century provided the final steps in mechanizing deduction sufficiently to enable the first automated theorem provers.

The modern offspring of this line of intellectual development include the many efforts that use first order logic as a representation and some variety of deduction as the reasoning engine, as well as the large body of work with the explicit agenda of making logical reasoning computational, exemplified by PROLOG.

This line of development illustrates clearly how approaches to representation are founded on and embed a view of the nature of intelligent reasoning. There is here, for example, the historical development of the underlying premise that reasoning intelligently *means* reasoning logically; anything else is a *mistake* or an abberation. Allied with this is the belief that "logically" in turn means first order logic, typically sound deduction. By simple transitivity these two collapse into one key part of the view of intelligent reasoning underlying logic: reasoning intelligently *means* reasoning in the fashion defined by first order logic.

A second important part of the view is the allied belief that intelligent reasoning is a process that can be captured in a formal description, particularly a formal description that is both precise and concise.

But very different views of the nature of intelligent reasoning are also possible. One distinctly different view is embedded in the part of AI influenced by the psychological tradition. That tradition, rooted in the work of Hebb, Bruner, Miller, and Newell & Simon, broke through the stimulus/response view demanded by behaviorism and suggested instead that human problem solving behavior could usefully be viewed in terms of goals, plans, and other complex mental structures. Modern manifestations include work on SOAR as a general mechanism for producing intelligent reasoning and knowledge based systems as a means of capturing human expert reasoning.

Comparing these two traditions reveals significant differences and illustrates the consequences of adopting one or the other view of intelligent reasoning. In the logicist tradition intelligent reasoning is taken to be a form of calculation, typically deduction in first order logic, while the tradition based in psychology takes as the defining characteristic of intelligent reasoning that it is a particular variety of human behavior. In the logicist view the object of interest is thus a construct definable in formal terms via mathematics, while for those influenced by the psychological tradition it is an empirical phenomenon from the natural world. There are thus two very different assumptions here about the essential nature of the fundamental phenomenon to be captured.

A second contrast arises in considering the character of the answers each seeks. The logicist view has traditionally sought compact and precise characterizations of intelligence, looking for the kind of characterizations encountered in mathematics (and at times in physics). The psychological tradition by contrast suggests that intelligence is not only a natural phenomenon, it is an inherently *complex* natural phenomenon: as human anatomy and physiology are inherently complex systems resulting from a long process of evolution, so perhaps is intelligence. As such, intelligence may be a large and *fundamentally* ad hoc collection of mechanisms and phenomena, one for which complete and concise descriptions may not be possible.

Several useful consequences result from understanding the different positions taken by each tradition on this fundamental question. First, it demonstrates that selecting any of the modern offspring of those traditions—i.e., any of the representation technologies shown at the bottom of the table—means choosing more than a representation: in the same act we are also selecting a conception of the fundamental nature of intelligent reasoning.

Second, those conceptions differ in important ways: there are fundamental differences in the conception of the phenomenon we are trying to capture. The different conceptions in turn mean there are deep-seated differences in the character and the goals of the various research efforts that are trying to create intelligent programs. Simply put: different conceptions of the nature of intelligent reasoning lead to different goals, definitions of success, and different artifacts being created.

Finally, those differences are rarely articulated. This in turn leads to arguments that may be phrased in terms of issues like representation choice (e.g., the virtues of sound reasoning in FOPC vs. the difficult to characterize inferences produced by frame-based systems), when the real issues are, we believe, the different conceptions of the fundamental nature of intelligence. Understanding the different positions assists in analyzing and sorting out the issues appropriately.

3.3.2 Which inferences are sanctioned?

The second component of a representation's theory of intelligent reasoning is its set of sanctioned inferences, i.e., a selected set of inferences deemed appropriate conclusions to draw from the information available. The classic definition is supplied by traditional formal logic, where the only sanctioned inferences are sound inferences (i.e., those encompassed by logical entailment, in which every model for the axiom set is also a model for the conclusion). This answer has a number of important benefits, including being satisfying intuitively (a sound argument will never introduce error), explicit (so we know precisely what we're talking about), precise enough that it can be the subject of formal proofs, and old enough that we have accumulated a significant body of experience with it.

Logic has also explored several varieties of unsound inference, including circumscription and abduction. This exploration has typically been guided by the requirement that there be "a well motivated model-theoretic justification" [19], as for example the minimal model criterion of circumscription. This requirement maintains a fundamental component of the logicist approach: while it is willing to arrive at conclusions that are true in some subset of the models (rather than true in every model) the set of sanctioned inferences is still conceived of in model theoretic terms and is specified precisely in those terms.

Other representations have explored other definitions: probabilistic reasoning systems (e.g., [20]) sanction the inferences specified by probability theory, while work on rational agents (e.g., [7]) relies on concepts from the theory of economic rationality.

Among the common knowledge representation technologies, rule based systems capture guesses of the sort the human expert makes, guesses that are not necessarily either sound or true in any model. A frame-based representation encourages jumping to possibly incorrect conclusions based on good matches, expectations, or defaults. Both of these share the psychological tradition of defining the set of sanctioned inferences with reference to the behavior of the human expert, rather than by reference to an abstract formal model.

3.3.2.1 Form and content of the answers

As these examples show, different approaches to representation specify sanctioned inferences in ways that differ in both content and form. Where the specification for logic, for example, is expressed in terms of model theory and is mathematically precise, other representations provide answers phrased in other terms, often with considerably less precision. Frames theory, for instance, offers a definition phrased in terms of human behavior and is specified only approximately.

The differences in both content and style in turn have their origin in the different conceptions of intelligent reasoning explored above. Phrasing the definition in terms of human behavior is appropriate for frames because the theory conceives of intelligent reasoning as a characteristic form of human behavior. In attempting to describe that behavior, the theory is faced with the task of characterizing a complex empirical phenomenon that can be captured only roughly at the moment, and that may never be specifiable with mathematical precision, hence the appropriateness of an approximate answer.

For frames theory, then, the specification of sanctioned inferences is both informal and empirical, as an unavoidable consequence of its conception of intelligence. The work (and other work like it) is neither sloppy nor causally lacking in precision; the underlying conception of intelligent reasoning dictates a different approach to the task, a different set of terms in which to express the answer, and a different focus for the answer.

The broader point here is to acknowledge the legitimacy of a variety of approaches to specifying sanctioned inferences: model theory may be familiar and powerful, but even for formal systems it is not the only possible language. More broadly still, formal definitions are not the only terms in which the answer can be specified. The choice of appropriate vocabulary and degree of formality depends in turn on the basic conception of intelligent behavior.

3.3.3 Which inferences are recommended?

While sanctioned inferences tell us what conclusions we are permitted to make, that set is invariably very large and hence provides insufficient constraint. Any automated system attempting to reason, guided only by knowing what inferences are sanctioned, will soon find itself overwhelmed by choices. Hence we need more than an indication of which inferences we can legally make, we also need some indication of which inferences which inferences are appropriate to make, i.e., intelligent. That indication is supplied by the set of recommended inferences.

Note that the need for a specification of recommended inferences means that in specifying a representation we also need to say something about how to reason intelligently. Representation and reasoning are inextricably, and usefully, intertwined: a knowledge representation *is* a theory of intelligent reasoning.

Reasonably often that theory has been provided by observation of human behavior. Minsky's original exposition of frame theory, for instance, offers a clear example of both a set of recommended inferences, and a set inspired by observation of human behavior. Consider this, from the introduction to the original frames paper [17]:

> This is a partial theory of thinking.... Whenever one encounters a new situation (or makes a substantial change in one's viewpoint), he selects from memory a structure called a *frame*; a remembered framework to be adapted to fit reality by changing details as necessary.
>
> A *frame* ... [represents] a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party.

The first sentence illustrates the intertwining of reasoning and representation: this is a paper about knowledge representation, yet it announces at the outset that it is also a theory of thinking. That theory in turn arose from an insight about human intelligent reasoning, namely how people might manage to make the sort of simple common sense inferences that appear difficult to capture in programs. The theory singles out a particular set of inferences to recommend, viz., reasoning in the style of anticipatory matching.

Similar characterizations of recommended inferences can be given for most other representation technologies. Semantic nets in their original form, for instance, recommend bi-directional propagation through the net, inspired by the interconnected character of word definitions and the part of human intelligence manifested in the ability of people to find connections between apparently disparate concepts. The rules in knowledge based systems recommend plausible inferences, inspired by observation of human expert reasoning.

Logic, by contrast, has traditionally taken a minimalist stance on this issue. The representation itself offers only a theory of sanctioned inferences, seeking to remain silent on the question of which inferences to recommend.

The silence on this issue is motivated by the desire for generality in the inference machinery and a declarative (i.e., use-independent) form for the language, both fundamental goals of the logicist approach: "...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." [19]

But a representation with these goals cannot single out any particular set of inferences to recommend, for two reasons. First, if the inference process is to be general and uniform (i.e., work on all problems, and work in the same way), it must be neutral about which inferences to recommend: any particular subset of inferences it attempted to single out might be appropriate in one situation but fatally bad in another, since no inference strategy (unit preference, set of support, etc.) is universally appropriate. Second, if statements in the language are to be declarative, they must express a fact without any indication of how to reason with it (use-free expression is a defining characteristic of a declarative representation). Hence the inference engine can't recommend any inferences (or it loses its generality and uniformity) and the statements of fact in the language cannot recommend any inferences (since by embedding such information they lose their their declarative character).[5]

The desire for generality and use-free expression thus prevents the representation itself from selecting inferences to recommend. But if the representation itself cannot make the recommendation, the user must, since the alternative—unguided search—is untenable.

This is in part a deliberate virtue of the logicist approach: preventing the representation from selecting inferences and hence requiring the user to do so offers the opportunity for that information to be represented explicitly, rather than embedded implicitly in the machinery of the representation (as, for example, in rule-based systems, or PROLOG.

One difficulty with this admirable goal arises in trying to provide the user with the tools to express the strategies and to guide the system. Three approaches are commonly employed: have the user tell the system what to do, have the user lead it into doing the right thing, and build in special-purpose inference strategies. By "telling the system what to do" we mean having the user recommend a set of inferences by writing statements in the same (declarative) language used to express facts about

[5]The consequences of this approach are evident even in the use of disjunctive normal form as a canonical representation: although $X1 \wedge X2 \wedge X3 \rightarrow X4$ is semantically equivalent to $\neg X1 \vee \neg X2 \vee X4 \vee \neg X3$, some potentially useful information is lost in the transformation. The first form might be read as suggesting that X1, X2, and X3 have something in common, viz., that they should be thought of as the preconditions needed to establish X4. This may be a useful hint about how to reason in the problem, but if so, it is lost in the transformation to DNF. Consider by contrast languages like microPLANNER and PROLOG, which make explicit use of the form of the inference rule to help guide the deduction process.

the world (e.g., MRS [9]). By "leading the system into doing the right thing" we mean having the user carefully select the axioms, theorems, and lemmas supplied to the system. The presence of a lemma, for instance, is not simply a fact the system should know, it also provides a way of abbreviating a long chain of deductions into a single step, in effect allowing the system to take a "large" step in a certain direction (viz., the direction in which the lemma takes us). With careful selection of facts and lemmas the user can indirectly recommend a particular set of inferences. By "special purpose inference strategies" we mean building specific control strategies directly into the theorem prover. This can offer significant speedup and a pragmatically useful level of computational efficiency.

Each of these approaches has both benefits and drawbacks. Expressing reasoning strategies in first order logic is in keeping with the spirit of the logicist approach, viz., explicit representation of knowledge in a uniform, declarative representation. But this is often problematic in practice: a language designed to express facts declaratively is not necessarily good for expressing the imperative information characteristic of a reasoning strategy.

Careful selection of lemmas is at best an indirect encoding of the guidance information to be supplied. Finally, special-purpose deduction mechanisms are powerful but embed the reasoning strategy both invisibly and procedurally, defeating the original goals of domain independent inference and explicit, declarative representation.

The good news here is that by remaining purposely silent on the issue of recommended inferences, logic offers both a degree of generality and the possibility of making information about recommended inferences explicit and available to be reasoned about in turn. The less good news is that the task of guiding the system has been left to the user with no conceptual assistance offered, and the practices that result at times defeat some of the key goals that motivated the approach at the outset.

3.4 Role IV: A KR is a Medium for Efficient Computation

From a purely mechanistic view, reasoning in machines (and somewhat more debatably, in people) is a computational process. Simply put, to use a representation we must compute with it. As a result, questions about computational efficiency are inevitably central to the notion of representation.

This has long been recognized, at least implicitly, by representation designers: along with their specification of a set of recommended inferences, representations typically offer a set of ideas about how to organize information in ways that facilitate making those inferences. A substantial part of the original frames notion, for instance, is concerned with just this sort of advice, as more of the introduction to the frames paper illustrates:

> A *frame* ... [represents] a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party.
>
> Attached to each frame are several kinds of information. Some of this

> information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed.

The notion of triggers and procedural attachment in frames is not so much a statement about what procedures to write (on this the theory is rather vague), as it is a description of a useful way to organize information: e.g., (paraphrasing from above) attach to each frame information about how to use the frame, what to do if expectations are not confirmed, etc. Similarly, organizing frames into taxonomic hierarchies both suggests taxonomic reasoning and facilitates its execution (as in structured inheritance networks).

Other representations provide similar guidance. Traditional semantic nets facilitate bi-directional propagation by the simple expedient of providing an appropriate set of links, while rule-based systems facilitate plausible inferences by supplying indices from goals to rules whose conclusion matches (for backward chaining) and from facts to rules whose premise matches (forward chaining).

While the issue of efficient use of representations has been addressed by representation designers, in the larger sense the field appears to have been historically ambivalent in its reaction. Early recognition of the notion of heuristic adequacy [16] demonstrates that early on researchers appreciated the significance of the computational properties of a representation, but the tone of much subsequent work in logic (e.g., [13]) suggested that epistemology (knowledge content) alone mattered, and defined computational efficiency out of the agenda. Epistemology does of course matter, and it may be useful to study it without the potentially distracting concerns about speed. But eventually we must compute with our representations, hence efficiency must be part of the agenda.

The pendulum later swung sharply over, to what we might call the computational imperative view. Some work in this vein (e.g., [15]) offered representation languages whose design was strongly driven by the desire to provide not only efficiency, but guaranteed efficiency. The result appears to be a language of significant speed but restricted expressive power [6].

Either end of this spectrum seems problematic: we ignore computational considerations at our peril, but we can also be overly concerned with them, producing representations that are fast but inadequate for real use.

3.5 Role V: A KR is a Medium of Human Expression

Finally, knowledge representations are also the means by which we express things about the world, the medium of expression and communication in which we tell the machine (and perhaps one another) about the world. This role for representations is inevitable so long as we need to tell the machine (or other people) about the world, and so long as we do so by creating and communicating representations.[6] The fifth role

[6]It will presumably continue to be useful even if machines invent their own knowledge representations based on independent experience of the world: if their representations become incomprehensible

for knowledge representations is thus as a medium of expression and communication for use by *us*.

That in turn presents two important sets of questions. One set is familiar: How well does the representation function as a medium of expression? How general is it? How precise? Does it provide expressive adequacy? etc.

An important question less often discussed is, How well does it function as a medium of communication? That is, how easy is it for *us* to "talk" or think in that language? What kinds of things are easily said in the language and what kinds of things are so difficult as to be pragmatically impossible?

Note that the questions here are of the form "how easy is it?" rather than "can we?" This is a language we must use, so things that are possible in principle are useful but insufficient; the real question is one of pragmatic utility. If the representation makes things possible but not easy, then as real users we may never know whether we have misunderstood the representation and just do not know how to use it, or it truly cannot express some things we would like to say. A representation is the language in which we communicate, hence we must be able to speak it without heroic effort.

4 Consequences For Research and Practice

We believe that this view of knowledge representation can usefully influence practice and can help inform the debate surrounding several issues in representation research. For practice it offers a framework that aids in making explicit the important insights and "spirit" of a representation, and illustrates the difference in design that results from indulging rather than violating that spirit.

The consequences of the view for research include (i) a broader conception of representation, urging that all of the roles should be kept in mind when creating representation languages, (ii) the recognition that a representation embeds a theory of intelligent reasoning, (iii) the ability to use the broader view of representation to guide the combination of representations, (iv) the ability to use the broader view to dissect some of the arguments about formal equivalence of representations; and (v) the belief that the central task of knowledge representation is capturing the complexity of the real world.

Space limitations require that we briefly sketch out these consequences here; a complete discussion is found in [5].

4.1 Consequence for Practice: Characterizing the Spirit of a Representation

The roles enumerated above help to characterize and make explicit the "spirit" of a representation, the important set of ideas and inspirations that lie behind (and, significantly, are often less obvious than) the concrete machinery used to implement the representation. The spirit is often difficult to describe with precision, but we

to us, the machines will be unable either to tell us what they know or explain their conclusions.

believe it is well characterized by the last four of the roles enumerated above (all representations are surrogates, hence there is little difference among them on the first role.) The stance a representation takes on each of these issues, along with its rationale for that stance, indicates what the representation is trying to say about how to view and reason about the world.

The frames idea, for example, is in its original incarnation [17] primarily an ontological commitment and a theory of intelligent reasoning based on insights about human cognition and the organization of knowledge in memory.

The major ontologic commitment of course is to viewing the world in terms of stereotypical descriptions, i.e., concepts described in terms of what is typically true about them. This is particularly well suited to concepts in the natural world, where categories rarely have precise specifications in terms of necessary and sufficient conditions, and where exceptions abound. There is additional ontologic commitment in linking frames into systems in order to capture perspective shifts; this encourages us to look for such shifts when viewing the world.

The theory of intelligent reasoning embedded in frames claims that much reasoning is based on recognition, particularly matching stereotypes against individual instances. The suggestions concerning organization of knowledge are based on the belief that information in human memory is richly and explicitly interconnected, rather than structured as a set of independent or only implicitly connected facts.

The frames theory thus recommends inferences produced by stereotype matching and instantiation, and facilitates those inferences by the frame structure itself, as well as the further organization of frames into frame systems.

The theory sanctions inferences that are unsound, as in the analogical and default reasoning done when matching frames. It also sanctions inferences that involve relatively large mismatches, in order to model understanding that is tenacious even in the face of inconsistencies.

The theory provides a medium for potentially efficient computation by casting understanding as matching rather than deduction. Finally, it offers a medium of expression that is particularly useful for describing concepts in the natural world, where we often need some way of indicating what properties an object typically has, without committing to statements about what is always true.

4.1.1 The spirit should be *indulged*, not *overcome*

Two useful consequences result from characterizing a representation in these terms, making its position on the roles both explicit and understandable. First, it enables a kind of explicitly invoked Whorfian theory of representation use. While the representation we select will have inevitable consequences for how we see and reason about the world, we can at least select it consciously and carefully, trying to find a pair of glasses appropriate for the task at hand. Steps in this direction would include having representation designers characterize carefully the nature of the glasses they are supplying (e.g., making explicit the ontologic commitments, recommended inferences, etc.) and having the field develop principles for matching representations to tasks.

Second, such characterizations would facilitate appropriate use of a representation. By appropriate we mean using it in its intended spirit, i.e., used for what it was *intended* to do, not for what it can be *made* to do. Yet with striking regularity, the original spirit of a representation is seen as an opponent to be overcome. With striking regularity the spirit gets forgotten, replaced by a far more mechanistic view that sees a data structure rather than a representation, computation rather than inference. Papers written in this mindset typically contain claims of how the author was able, through a creative, heroic, and often obscure act, to get a representation to do something we wouldn't ordinarily have thought capable of doing.

But if that's what using a representation is all about, it becomes an odd and terribly awkward form of programming: the task becomes one of doing what we already know we want to do, but being forced to do it using just the constructs available in the representation, no matter how good or bad the fit.

In that case the creative work is often in *overcoming* the representation, seeing how we can *get it to behave like something else.*[7] The result is knowledge representations applied in ways that are uninformed by the inspirations and insights that led to their invention and that are the source of their power. Systems built in this spirit often work despite their representations, not because of them; they work because the authors, though great effort, managed to overcome the representation.

4.2 Consequence for Research: Representation and Reasoning are Intertwined

At various times in the development of the field, the suggestion has been made that we ought to view knowledge representation in purely epistemologic terms, i.e., take the singular role of representation to be conveying knowledge content (e.g., [13]). As we noted earlier, epistemology matters, but it is not the whole of the matter. Representation and reasoning are inextricably intertwined: we cannot talk about one without also, unavoidably, discussing the other. We argue as well that the attempt to deal with representation as knowledge content alone leads to an incomplete conception of the task of building an intelligent reasoner.

Each of these claims is grounded in an observation made earlier. We observed first that every representation embeds at its core a conception of what constitutes intelligent reasoning (e.g., Table I). Hence any discussion of representation unavoidably carries along with it a (perhaps implicit) view of intelligent reasoning.

We also observed that, in building an intelligent reasoner it is not enough to indicate what inferences are legal, we also need to know which are appropriate (i.e., recommended). A familiar example from logic makes the point nicely: from `A` we can

[7] There is of course utility in establishing the equivalence of two representations by showing how one can be made to behave like another. But this is an exercise that need be done only once, and it is done for its own sake rather than because it is good system construction practice.

infer $\mathtt{A} \wedge \mathtt{A}$, $\mathtt{A} \wedge \mathtt{A} \wedge \mathtt{A}$, etc. All of these are *legal* inferences, but they are hardly *intelligent* inferences.

Hence our claim that a theory of legal (sanctioned) inference is thus insufficient; to build an intelligent reasoner we also need a theory of intelligent inference. There may in fact be multiple theories of intelligent inference, each specific to a particular task domain.

4.3 Consequence for Research: Combining Representations

Recall that a representation is among other things a theory of intelligent reasoning and a collection of mechanisms for implementing that theory. We believe that appropriate attention to both of these aspects, in the appropriate order, makes a significant difference in the outcome of any effort at representation combination.

Too often, efforts at combination appear to be conceived of in terms of finding ways for the two *mechanisms* to work together, with insufficient (and sometimes no) attention to what we consider to be the much more important task: determining how the two theories of intelligent reasoning might work together. Focusing on mechanisms means determining such things as how, say, rules, procedures, and objects might invoke one another interchangeably. Focusing on the theories of intelligent reasoning means attending to what kinds of reasoning is within the spirit of each representation and how those varieties of reasoning might sensibly be combined.

Two efforts at combining representations illustrate the different conceptions of the task that arise from focusing on mechanism and from focusing on reasoning. As the first example, consider, this description of the LOOPS system [22] and its efforts at integrating several paradigms:

> Some examples illustrate the integration of paradigms in LOOPS: the "workspace" of a ruleset is an object, rules are objects, and so are rulesets. Methods in classes can be either LISP functions or rulesets. The procedures in active values can be LISP functions, rulesets, or calls on methods. The ring in the LOOPS logo reflects the fact that LOOPS not only contains the different paradigms, but integrates them. The paradigms are designed not only to complement each other, but also to be used together in combination.

Contrast the mindset and approach above with this view of a similar undertaking, also aimed at combining rules and frames [23]:

> Rules and frames are two contrasting schemes for representing different kinds of knowledge. Rules are appropriate for representing logical implications, or for associating actions with conditions under which the actions should be taken. .... Frames (or semantic nets) are appropriate for defining terms and for describing objects and the taxonomic

> class/membership relationships among them. An important reasoning capability of frame systems with well-defined semantics is that they can infer the class/membership relations between frames based on their definitions.
>
> Since the strengths and weaknesses of rules and frames are complementary to each other, a system that integrates the two will benefit from the advantages of both techniques. This paper describes a hybrid architecture called *classification based programming* which extends the production system architecture using automatic classification capabilities within frame representations. In doing so, the system enhances the power of a pattern matcher in a production system from *symbolic matching* to *semantic matching*, organizes rules into rule classes based on their functionalities, and infers the various relationships among rules that facilitate explicit representation of control knowledge.

Note in particular how the first of these efforts focuses on computational mechanisms, while the second is concerned with representation and reasoning. The first seeks to allow several programming constructs—among them rules and structured objects—to work together, while the second attempts to allow two representations—rules and structured objects—to work together. The first seeks to permit mechanisms to invoke one another, while the second considers the different varieties of inference natural to two representations and suggests how those two kinds of reasoning could work synergistically: rules are to be used for the kind of reasoning they capture best—unrestricted logical implications—and frames are to be used for their strength, viz., taxonomic reasoning. The first paper thus proposes a *computational architecture*, while the second offers a *representation and reasoning architecture*.

Both of these are important undertakings, but we believe that where the goal is combining representations, the task should be conceived of in terms central to a representation: its theory of intelligent reasoning. To do this we should consider what kind of reasoning we expect from each representation, we should propose a design for how those reasoning schemes will work together, and we should have a rationale for believing that this will be an effective combination. The first and most important thing we require in this task is a representation architecture; from that the appropriate computational architecture will follow. We believe that efforts that attend to only the second of these are both beginning the effort in the wrong place and will fall far short of a crucial part of the goal.

4.4 Consequence for Research: Arguments About Formal Equivalence

There is a familiar pattern in knowledge representation research in which the description of a new knowledge representation technology is followed by claims that the new ideas are in fact formally equivalent to an existing technology. Historically the claim has often been phrased in terms of equivalence to logic. Semantic nets, for example, have been described in these terms (e.g., [11], [18]), while the development of frames led to a sequence of such claims, including the suggestion that "Most of 'frames' is *just* a new syntax for first-order logic" ([13], emphasis added).

That frames may *also* be an alternative syntax seems clear; that they are *merely* or *just* that seems considerably less clear. That frames are not entirely distinct from logic seems clear; that all of the idea can be seen as logic seems considerably less clear.

We believe that claims like those above are substantive only in the context of a narrowly construed notion of what a knowledge representation is. Hayes [13] is explicit about part of his position on this: "One can characterise a representational language as one which has (or can be given) a semantic theory" and explicit about the tight lines drawn around the argument: "Although frames are sometimes understood at the metaphysical level and sometimes at the computational level, I will discuss them as a representational proposal," i.e., as a language with a semantic theory, and nothing more. Both metaphysics and computation have been defined as out of the agenda. And again: "None of this discussion [about frames as a "computational device" for organizing memory access and inference] engages representational issues." Here it becomes evident that the claim is less about frames and more a definition of what will be taken as constituting "representational issues."

Note that specifically excluded from the discussion are the ontological commitment of the representation, viz., "what entities shall be assumed to exist in the world," and the computational properties the representation provides. Despite claims to the contrary, we would argue that the ontology of frames (and other representations) and computational questions not only *engage* representational issues, they *are* representational issues. These and other properties are crucial to knowledge representation both in principle and in any real use.

4.5 Consequence for Research: All Five Roles Matter

While representations are often designed with considerable attention to one or another of the issues listed in our five roles, we believe that all of the roles are of substantial significance and that ignoring any one of them can lead to important inadequacies. While designers rarely overlook representation's role as a surrogate and as a definition of sanctioned inferences, insufficient guidance on the others roles is not uncommon, and the consequences matter.

As we have argued above, for instance, ontologic commitment matters: the guidance it provides in making sense of the profusion of detail in the world is among the most important things a representation supply. Yet some representations, in a quest for generality, offer little support on this dimension.

A similar argument applies to the theory of intelligent reasoning: a representation can guide and facilitate reasoning if it has at its heart a theory of what reasoning to do. Insufficient guidance here will leave us susceptible to the traditional difficulties of unguided choice.

Pragmatically efficient computation matters because most of the use of a representation is (by definition) in the average case. Interest in producing weaker representations in order to guarantee improved worst-case performance may be misguided, demanding far more than is necessary and paying a heavy price for it.[8]

The use of a representation as a medium of expression and communication matters because we must be able to speak the language in order to use it. If we can't determine how to say what we're thinking, we can't use the representation to communicate with the reasoning system.

Attempting to design representations to accommodate all five roles is of course challenging, but we believe the alternative is the creation of tools with significant deficiencies.

5 The Goal of KR Research

We believe that the driving preoccupation of the field of knowledge representation should be understanding and describing the richness of the world. Yet in practice research that describes itself as core knowledge representation work has concentrated nearly all of its efforts in a much narrower channel, much of it centered around taxonomic and default reasoning (e.g., [2], [15], [8]).

We believe it is not an accident that useful insight about finding a good set of temporal abstractions came from close examination of a realistic task set in a real-world domain. It underscores our conviction (shared by others; see [14]) that attempting to describe the richness of the natural world is the appropriate forcing function for KR work

Our point here concerns both labeling and methodology: (i) work like [10] and [14] should be recognized by the KR community as of central relevance to KR research, not categorized as diagnosis or qualitative physics and seen as unrelated, and (ii) insights of the sort obtained in [10] and [14] come from studying the world, not from studying languages. We argue that those who choose to identify themselves as KR researchers should be developing theory and technology that facilitates projects like these; and conversely those who are building projects like these are engaged in a centrally important variety of KR research.

While tools and techniques are important, the field is and ought to be much richer than that, primarily because the world is much richer than that. We believe that understanding and describing that richness should be the central preoccupation of the field.

6 Summary

We have argued that a knowledge representation plays five distinct roles, each important to the nature of representation and its basic tasks. Those roles create multiple,

[8] As we have argued elsewhere [4], a "computational cliff" (i.e., unacceptable worst case behavior) is a problem only if every inference, once set in motion, cannot possibly be interrupted. The simple expedient of resource limited computation prevents any inference from permanently trapping the program in a task requiring unreasonable amounts of time.

sometimes competing demands, requiring selective and intelligent tradeoff among the desired characteristics. Those five roles also aid in characterizing clearly the spirit of representations and representation technologies that have been developed.

This view has consequences for both research and practice in the field. On the research front it argues for a conception of representation broader than the one often used, urging that all of the five aspects are essential representation issues. It argues that the ontological commitment a representation supplies is one of its most significant contributions; hence the commitment should be both substantial and carefully chosen. It also suggests that the fundamental task of representation is describing the natural world and claims that the field would advance furthest by taking this as its central preoccupation.

For the practice of KR work, the view suggests that combining representations is a task that should be driven by insights about how to combine their theories of intelligent reasoning, not their implementation mechanisms. The view also urges the understanding of and indulgence of the fundamental spirit of representations. We suggest that representation technologies should not be considered as opponents to be overcome, forced to behave in a particular way, but should instead be understood on their own terms and used in ways that rely on the insights that were their original inspiration and source of power.

Acknowledgments

Earlier versions of this material formed the substance of talks delivered at the Stanford Knowledge Systems Lab and an invited talk at AAAI–91; many useful comments from the audience were received on both occasions. Jon Doyle and Ramesh Patil offered a number of insightful suggestions on drafts of this paper; we are also grateful to Rich Fikes, Pat Hayes, Ron Brachman, and Hector Levesque for extended discussions of some of this material.

References

[1] Brachman R, Levesque H, eds., *Readings in Knowledge Representation*, Los Altos: Morgan Kaufman. 1985.

[2] Brachman, R. et. al. KL-ONE Reference Manual, BBN Report No. 3848, July 1978, Bolt, Beranek and Newman, Cambridge, MA.

[3] Davis R, Shrobe H E, Representing Structure and Behavior of Digital Hardware, *IEEE Computer*, Special Issue on Knowledge Representation, 16(10):75-82.

[4] Davis R, A Tale of Two Knowledge Servers, *AI Magazine*, **12**(3):118–120, Fall 1991.

[5] Davis R, Shrobe H, Szolovits P, What is a Knowledge Representation, Memo, MIT AI Lab, forthcoming.

[6] Doyle J, Patil R, Two dogmas of knowledge representation, MIT LCS Technical Memo 387B, September 1989. A revised version of this paper appeared as Doyle J., and Patil R, Two theses of knowledge representation: Language restrictions, taxonomic classification, and the utility of representation services, *Artificial Intelligence*, **48**(3):261–297, 1991.

[7] Doyle J., Rationality and Its Roles in Reasoning, *Computational Intelligence*, **8**:376–409, 1992.

[8] Fahlman S, Touretzky D and van Roggen W, Cancellation in a Parallel Semantic Network, In *Proc IJCAI-81*, 257–263, Menlo Park: International Joint Conferences on AI.

[9] Russell S, The Compleat Guide to MRS, Report No. STAN–CS–85–1080, Stanford Computer Science Dept, June 1985 (see esp. chapter 6).

[10] Hamscher W, Modeling digital circuits for troubleshooting, *Artificial Intelligence*, **51**:223–272, October 1991.

[11] Hayes P, In defense of logic, in *Proc. IJCAI-77*, 559–565, Menlo Park: International Joint Conferences on AI.

[12] Hayes P, Naive physics I: Ontology for liquids. University of Essex report, 1978, Essex, UK.

[13] Hayes P, The Logic of Frames, reprinted in *Readings in Knowledge Representation*, pp. 288-295.

[14] Lenat D, et al., CYC: Towards programs with common sense, *CACM*, **33**(8):30–49.

[15] Levesque H, Brachman R, A fundamental tradeoff in knowledge representation and reasoning, reprinted in *Readings in Knowledge Representation*, pp. 42–70.

[16] McCarthy J, Hayes P, Some philosophical problems from the standpoint of AI, in *Machine Intelligence 4*, Meltzer and Michie (eds.), Edinburgh: Edinburgh University Press, pp. 463-504, 1969.

[17] Minsky M, A framework for representing knowledge, Memo 306, MIT AI Lab, June 1974. (A revised version appeared as: Minsky, M., A framework for representing knowledge, Chapter 6 in *The Psychology of Computer Vision*, Winston, P. H. (ed.), McGraw-Hill, 1975.)

[18] Nash-Webber B, Reiter R, Anaphora and logical form, in: *Proc IJCAI-77*, 121–131, Menlo Park: International Joint Conferences on AI.

[19] Nilsson N, Logic and artificial intelligence, *AI*, **47**:31–56, January 1991.

[20] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Los Altos: Morgan-Kaufmann, 1988.

[21] Pople H, Heuristic methods for imposing structure on ill-structured problems, in *AI in Medicine*, Szolovits (ed.), AAAS Symposium 51, Boulder: Westview Press.

[22] Stefik M, Bobrow D, Mittal S, and L Conway, Knowledge programming in LOOPS: Report on an experimental course, *AI Magazine*, 4(3):3–10.

[23] Yen J, Neches R, MacGregor R, Classification-based Programming: A deep integration of frames and rules, Report ISI/RR–88–213, USC, Information Sciences Institute, Marina del Rey, CA, April 1989.