Systems Research Vol. 6, No. 1, pp. 39–52, 1989.
Printed in Great Britain.

0731–7239/89 $3.00+0.00
Pergamon Press plc

Artificial Intelligence and Philosophy: The Knowledge of Representation

MARCELO DASCAL

Tel Aviv University, Israel

(*Received* 14 *March* 1988; *in revised form* 6 *October* 1988)

Key words—Artificial intelligence; representation; knowledge; epistemology; expert systems; justification; pragmatics; imputation; philosophy; inference, defeasible rules; context-dependency.

Abstract—Philosophy has recently witnessed a radical critique of the epistemological tradition centered on the notion of representation. It is here argued that such a critique is relevant to the current endeavors of Artificial Intelligence, both in its broad (i.e. as an attempt to elucidate human cognitive capacities) and narrow (i.e. as a practical concern to develop 'intelligent' systems, e.g. expert systems) definitions. It is shown that many of the recurrent problems faced by AI researchers stem from their espousal of the old paradigm conceptions of 'knowledge', 'understanding', 'justification', and the like. Some suggestions concerning a possible alternative for such conceptions are made.

INTRODUCTION

THAT PHILOSOPHY is relevant to AI is hardly a thesis that needs defending. From the early days of AI, researchers have turned to philosophy for tools and insights. Formal logic and formal semantics have been instrumental in the development of AI. So too have philosophy of science, philosophy of mind, philosophy of language and epistemology. Even criticism of AI by philosophers who attempted to legislate on what computers can and cannot do, though discarded as irrelevant by many AI researchers, has forced some to think more thoroughly about the prospects and limitations of their enterprise. Witness the fact that 'cognitive science', the relatively new discipline that pursues many of the aims of AI, is usually conceived as an interdisciplinary endeavor, of which philosophy is an integral and undisputed part.

What I want to argue—and what needs to be argued for—is that the relevance of philosophy to AI has been conceived so far in a rather narrow way, and that, as a result, some implications of such a relevance have been overlooked.

The notions of 'representation' and 'knowledge' have occupied and still do occupy a focal position in the latest developments in Artificial Intelligence. Thus, for example, data banks are said to contain 'knowledge' whose 'representation' in a suitable format is supposed to enable 'expert systems' to use it for the purposes of their designers and users. At the level of design of such systems, discussion for the most part centers on the characterization of the best format for representing knowledge in a given field, or on the means to extract from human experts the knowledge they have and transfer it to the system. At the theoretical level, discussion focuses on such problems as the amount and kinds of knowledge needed by an 'intelligent' system (e.g. should it include 'common sense' knowledge or only knowledge specific to the field of expertise?), the kind of 'logic' or 'inference machine' required by the system (e.g. general, formal logic, or 'special-purpose' logic?), and what has been and what could be achieved by such systems. But the concepts of 'knowledge' and 'representation' themselves are usually taken as non-problematic, and seldom are the subject of critical scrutiny. To be sure, there is increasing dissatisfaction with the current ways of conceptualizing the problems in AI, and there is a search for 'heterodox' views. Many researchers are aware of developments in neurophysiology and cognitive psychology that question both the soundness of the traditional concepts of knowledge and representation and the need to use them as building blocks of a theory of human cognition. But few are aware of the *philosophical* critique to which such concepts have been recently suggested. My main purpose here is to draw attention to this critique and to its implications for AI. It seems to me that the philosophically-based critique of 'knowledge' and 'representation', and the alternatives it suggests, contrast significantly with the neurophysiologically—and cognitive—psychologically-based critiques. For both in fact suggest that, once representations are dissolved as a respectable theoretical construct or at least removed from their focal position in an account of mental life, one is free to propose explanations of what really goes on in the 'mind' (or in a mind-like system) in terms of more basic characteristics of the *individual*, be they holistic or fuzzy neurological patterns or else diffuse, pre-categorial, mental 'presentations'. The upshot of the philosophical critique of representations I am about to describe, on the other hand, is that it is not by digging deeper into the individual's head that one discovers the relevant parameters of his mental life. For these parameters are in fact social, not individual, public, not private, context-relative, not universal. In thus linking directly the cognitive with the social, and applying the results to disclose some intrinsic weaknesses of Artificial Intelligence's use of the notions of knowledge and representation, I believe the philosophical approach here described suggests a different kind of heterodox alternative to the prevalent views on the relationship between minds, societies, and machines.

I will first (Section 1) recall that AI's 'traditional' philosophical partner, namely analytic philosophy, has recently come under severe criticism, not only from competing philosophical schools, but also from within. As a result, it can no longer be taken for granted that 'interacting with philosophy' is reducible to 'interacting with analytic philosophy'. Both alternative approaches to philosophizing and the consequences of the critique for certain key conceptions characteristic of analytic philosophy and the tradition it derives from must be taken into account by any discipline—including AI—that purports to interact with philosophy. My next step (Section 2) will be to summarize the main points of Richard Rorty's critique of the crucial traditional and analytic notions of knowledge and representation. I will then ask whether such a critique has any impact on AI's handling of these and related notions. At first (Section 3), this will be discussed with respect to a 'broad' notion of AI, namely its conception as a discipline that endeavors to use computer programs 'as tools in the study of intelligent processes, tools that help in the discovery of the thinking-procedures and epistemological structures employed by intelligent creatures' [5, p. 17]. Then (Section 4), I will consider the impact of the above mentioned critique upon a narrower conception of AI, namely as a technical endeavor to create programs that perform intelligent tasks for specific purposes. The conclusion, in both cases, will be that the critique is highly relevant, and that, unless AI dissociates itself from some of the assumptions it unreflectively borrowed from epistemology, it will be unable to cope with some of the problems it now faces. Finally (Section 5) I will present an alternative, 'pragmatic', notion of knowledge, in the light of which the problem of 'representation of knowledge' looks much less urgent and important for AI than it is usually taken to be.

SECTION 1

Artificial Intelligence research has been very selective in its recourse to philosophy. Given its interest and purposes, this was only too natural. When it had to come to grips with the implementation of inferential systems, where to look for inspiration if not in formal logic? When concerned with language-understanding systems, what was more natural than to consult philosophy of language and formal semantics? And when interested in the issue of knowledge representation, what could be a better philosophical partner than the theory of knowledge (epistemology) and the philosophy of science?

However, in appealing to these well entrenched and well developed philosophical disciplines, AI research was, *ipso facto*, establishing an implicit pact with a certain historically specific way of philosophizing, namely 'analytic philosophy', and within it, with a small number of sub-disciplines (which excluded, among others, ethics, aesthetics, metaphysics, and the philosophy of action). More importantly, such a pact ruled out any significant interaction with alternative approaches to philosophizing—roughly those approaches, bundled together (by analytical philosophers) under the label 'continental philosophy', which analytic philosophy itself considered to be 'unphilosophical' because they did not live up to *its* standards of clarity and rationality. This includes, among other things, phenomenology, existentialism, hermeneutics, historicism, critical social theory, and marxism. It might be argued that the exclusion of such potential partners was in fact not the choice of AI (nor, for that matter, of analytic philosophy) but the other way around. After all, most of these approaches are grounded on a more or less sweeping criticism of the mode of rationality embodied by 'science', which they consider to be excessively narrow, and whose imperialism they denounce as a danger to a truly human conception of man and society. They tend to consider AI as epitomizing this narrow conception of rationality, which deprives the notions of 'thinking' and 'intelligence' of their true and deep human meaning. In addition to their wholesale rejection of AI, criticisms stemming from the above mentioned approaches were in many cases appallingly ill-informed or at least naive in their claims about how computer programs actually work. In the light of such claims and attitudes, no wonder that most AI researchers felt that there was nothing to talk about with or to learn from 'continental philosophy'.

Analytic philosophy, on the other hand, even when critical, was much more congenial. It shared with AI researchers both a methodological commitment to explicitness, to the value of formalization, to limited but clearly defined goals, to the Cartesian principle of analysis (i.e. breaking down a complex problem into its parts and combining the partial solutions into a global one), and a concern for roughly similar substantive issues such as the specification of the relationship(s) between knowledge and the world, the elucidation of the nature of cognitive processes,and a special emphasis on the role of language therein.

This symbiosis of AI with analytic philosophy was a satisfactory (though somewhat restrictive) way of fulfilling AI's need for a philosophical partner as long as the analytic tendency was able to hold the dominant position it had progressively conquered since the 1920s, at least in the anglo-American philosophical scene, by successfully setting aside both 'continental' criticism and internal dissent as either 'external' or 'marginal'. Yet, this situation has changed in recent years. A powerful current of internal criticism of analytic philosophy and of the whole tradition it descends from, by philosophers who were themselves analytically trained (such as Richard Rorty), has led to a deep questioning of its most basic methodological and substantive tenets. This same movement has led to a reappraisal of the significance of the earlier rejected 'continental' approaches. As a consequence, the dominance of the analytic paradigm in philosophy has been challenged, if not altogether overrun, and there is a hectic search for a new, 'post-analytic' mode of philosophizing. Thus, at this moment, it can no longer be taken for granted that 'interacting with philosophy' means 'interacting with analytic philosophy'.

SECTION 2

Though such a criticism stems from different sources and has been voiced by many authors, I will focus here on the version that argues that the reliance on the notion of 'representation' is the main reason for the failure of traditional epistemology.*

The main culprit in Rorty's [29] critique of the philosophical tradition that begins in the 17th century and goes on unchallenged up to 20th century's analytic philosophy is the view that the mind is 'a great mirror, containing various representations—some accurate, some not' [29, p. 12]. On this view, knowledge is accuracy of representation, and the task of 'epistemology' is to tell us more about the 'foundation of knowledge' by inspecting (and eventually polishing and repairing) the mirror. Descartes', Locke's and Kant's methods, as well as analytic philosophy's 'conceptual analysis' and 'linguistic analysis' only make sense if the correctness of that view is assumed. So too do the hopes to replace philosophical epistemology by such possible successor disciplines as empirical psychology (or cognitive science) and philosophy of language, as illustrated for instance by Quine's 'epistemology naturalized' or by his method of 'semantic ascent', or else by Piaget's 'genetic epistemology'. According to Rorty, the trouble with all these proposals is that they think of knowledge as presenting 'a problem', about which we ought to have 'a theory', and this, in turn, 'is a product of viewing knowledge as an assemblage of representations' [29, p. 136]. On this view, a theory of knowledge is a theory of representation(s), whose emphasis may lie in establishing privileged, uncorrigible representations (Descartes), or in describing the machinery which enables us to build complex representations out of simple ones (Locke), or else in resolving the problem of how 'two irreducibly distinct sorts of representations—"formal" ones (concepts) and "material" ones (intuitions)' [29, p. 138]—are related (Kant). Such a theory is supposed to provide an *a priori* account of both the 'foundations' of all knowledge and of the limits of our knowing ability, and thus to solve the problem of knowledge by providing the principles that specify the legitimate range of operation of our cognitive apparatus when applied to any extant or possible subject matter. On Rorty's view, all 'theories of knowledge', classical or contemporary, are animated by a desire to determine universal, invariant rules governing the 'rational' or 'scientific' uses of our understanding; they share 'the urge to see social practices of justification as more than just such practices' [29, p. 390].

Among the many specific points that can be derived from Rorty's attempt to 'deconstruct' the deeply entrenched ocular metaphor that describes the mind as a mirror-like assemblage of representations, the following are particularly relevant to our concerns here:

(i) The conception of the 'mind' as a distinct, self-contained sphere of inquiry, was an invention of 17th century philosophers. There is nothing necessary nor intuitive about it. Consequently, there need not be any special distinctive characteristics of this sphere, such as 'intentionality', 'representationality' and 'meaning'. Such alleged features of 'the mental' are not extra 'immaterial' properties that familiar physical objects (such as inscriptions) have, but simply 'the part they play in a larger context—an interaction with large numbers of other visible things' [29, p. 27]. Here we are catering on the contributions of such philosophers as Ryle and Wittgenstein who attacked the (Cartesian) myth of 'the ghost in the machine'.

* Richard Rorty, whose work will be our main source, is not the only one to criticize the use of the notion of representation in epistemology. A similar criticism, albeit stemming from a different outlook, can be found in Taylor [33]. In the non-analytic camp, many have criticized this notion before Rorty. What is characteristic of Rorty's contribution is his combined use of analytical and non-analytical sources for his critical purposes.

On the face of it, this seems to support the view that AI *can* provide a model of mental processes, against critics who argue that the mental has certain exclusive characteristics, such that no computer program, sophisticated as it may be, cannot in principle possess. An example of this kind of criticism is John Searle's [31] famous comparison of a computer program with the manipulation of Chinese characters by someone who does it quite well but has no idea of their meanings. According to Searle, the exclusive human (mental) characteristic that the program (and the manipulator of the characters, in this case) does not have is Intentionality (one of whose instances is precisely the 'meaning' which is absent from the performance of the character manipulator). To be sure, Rorty's argument—if correct—is ruinous for Searle's contention. But, alas, it is of no relief for AI. For Rorty's attack is intended to *dissolve* the sphere of 'the mental' as something that requires a special account or explanation, so that, in so far as AI purports to model, imitate, represent, or provide a theory of 'the mental', it would become jobless, if the attack were to succeed.

(ii) Since the 'mind' is not a separate sphere, the attempt to provide 'a para-mechanical account of mental processes' or a 'physiology of the human understanding' is pointless. In particular, such an account is not—by any means—to be obtained from an inspection of recondite 'inner mechanisms', whether they lie in the human brain or in computer programs. The 'mental', in so far as it makes sense to single it out, is not private but public: the 'interaction with other visible things' referred to above is mainly social interaction between (human) agents. It is within the global context of such an interaction that meaning and other kinds of intentionality emerge, and it is only by reference to such a context that these allegedly 'mental' phenomena can be understood and accounted for in a non-mysterious way.

(iii) The belief that some equivalent of a 'physiology of the human understanding', if available, would provide a solution to the 'problem of knowledge', by way of offering a criterion to identify the accuracy of those representations entitled to be called 'knowledge', is based on a confusion between causal explanation and justification. This is what Rorty describes as 'Locke's mistake'. It consists in conflating two questions, namely (a) what 'counts as' knowledge and (b) under what conditions does a (human) organism come to have knowledge, and assuming that the answer to the latter is also an answer to the former. This amounts to supposing that the 'logical space of reasons' can be reduced to the 'logical space of causes', that epistemic notions such as belief and knowledge can be analyzed without remainder into non-epistemic facts. This thesis, later dubbed 'psychologism', was strongly criticized by philosophers such as Frege and Husserl. But it could be argued that such criticism did not touch its essential assumption, which is the reduction of 'knowledge that such and such is the case' to 'knowledge of' a certain (mental) object which we 'have in mind'. On this view, it is either the intrinsic nature of this object (Descartes) or its causal story (Locke) that ensures its status as 'knowledge'. The alternative is to consider knowledge not as a relation to objects but as a matter of justification of certain beliefs, i.e. as belonging to an entirely different game: not 'inspection' but 'argumentation'.

(iv) If one makes 'Locke's mistake', itself a result of uncritically adopting the ocular metaphor, one is compulsively led to the belief that knowledge needs 'foundations', to be discovered by turning the Mind's Eye to the appropriate representations in the mind:

> . . . we will want to get behind reasons to causes, beyond argument to compulsion from the object known, to a situation in which argument would be not just silly but impossible, for anyone gripped by the object in the required way will be *unable* to doubt or see an alternative. To reach that point is to reach the foundations of knowledge. For Plato, that point was reached by escaping from the senses and opening up the faculty of reason—the Eye of the Soul—to the World of Being. For Descartes, it was a matter of turning the Eye of the Mind from the confused inner representations to the clear and distinct ones. With Locke, it was a matter of reversing Descartes's directions and seeing 'singular presentations to sense' as what should 'grip' us—what we cannot and should not wish to escape from. [29, p. 159]

(v) This search for 'foundations' presupposes, in addition, (a) that it is possible to distinguish, atomistically, between different kinds of representations, and (b) that some of them (e.g. 'sense-data', 'protocol sentences', etc.) are somehow 'epistemically privileged'. But Quine's holistic criticism of the analytic–synthetic distinction, and Sellars's attack on the 'myth of the given' show that both assumptions are untenable. Such criticisms, coupled with Wittgenstein's emphasis of the public, social nature of language and Davidson's demolition of the scheme/content distinction, 'successfully, and rightly, blur the positivist distinctions between the semantic and the pragmatic, the analytic and the synthetic, the linguistic and the empirical, theory and observation'. Analytic philosophy, which culminates in the work of the four thinkers mentioned, thus 'transcends and cancels itself' [30]. Analytic philosophy 'cancels itself' because it simply suppresses the traditional conception of Philosophy—with a capital P—as the 'queen of sciences', as the privileged discipline that establishes universal criteria for what is to count as knowledge and what not, and describes the method whereby knowledge is to be achieved.

(vi) An alternative to the criticized doctrine is to view knowledge or 'rational certainty' not as a special relation to an object known, but rather as a matter of what we are justified in believing. This, in turn, is related to the social phenomenon of discussing with others, arguing for a thesis, and eventually winning the argument, i.e. it is a function of our ability to countenance objections and to persuade our interlocutors, rather than of our relation—as 'knowing subjects'—to 'reality' [30, p. 9]. On this view, it is not by looking into our faculties that we will discover more about knowing, believing, etc., but rather by looking into our social interactions with our fellow humans. Since 'victory in argument'—which is all there is to 'justification'—is contingent upon a host of unpredictably variable (contextual) factors, there is no absolute 'foundation' of knowledge or truth, no privileged 'neutral' representation medium or 'conceptual scheme', no 'method' that would enable an 'impartial judge of controversies' to decide 'objectively' on any given knowledge claim. In so far as 'epistemology' is, by its very nature, the search for such utopic contrivances, it should simply cease to exist. This should also be the destiny of any discipline—such as cognitive science—that purports to succeed where epistemology has failed.

SECTION 3

In spite of its devastating consequences for a certain conception of knowledge—which I think is the one underlying the use of this term in AI—not everything in Rorty's criticism is bad news for AI. After all, his concern is strictly internal to philosophy. All he wants to achieve—as we have seen—is a deconstruction of a certain conception of this discipline as an overall arbiter, above the sciences and other social practices, of knowledge claims. In so far as AI could be conceived as just one of these social practices, a 'form of discourse' within which no absolute truth or knowledge claims are made, but merely claims that can be judged in terms of 'the successful accomplishment of (that) practice' [29, p. 319], it could presumably be immune to those criticisms. We shall return to this possibility in the next section. But we must first deal with AI in its more ambitious form, which is not totaly absent—in spite of appearances—from its more modest manifestations.

In many respects, the 'broad' conception of AI (AI[B]) as a discipline that purports not only to relieve us from performing boring intellectual tasks and to help us in dealing with tasks whose complexity depass our limited computational capacity, but also to enable us to learn more about our own 'mental life', inherits both the ends and the assumptions of traditional 'epistemology'. In so far as this is the case, AI, broadly conceived, is subject to essentially similar criticisms, with some specific additions and modifications.*

(i) AI[B], just like traditional epistemology, seeks to discover the nature of our cognitive (as well as other mental) processes by developing systems that are able to emulate (some of) our cognitive (and other mental) performances. Such systems are to be viewed as theoretical hypotheses about our cognitive processes, which can be assessed on the basis of general epistemological criteria (e.g. descriptive adequacy, explanatory adequacy, simplicity, fecundity, etc.). In so far as they withstand such tests, these systems can be considered to be 'models' of the mental processes in questions. Thus, they teach us about these processes because they stand in a relation of *adequate representation* to them. It is by inspecting these representations or models *of* our cognitive processes, that we are supposed to gain knowledge *that* such processes have this or that property. AI[B] thus adopts the traditional view that knowledge-that can be derived from (and justified by) knowledge-of.

This assumption underlies AI's fondness for the phrase 'representation of knowledge', which should in fact be read the other way around, namely 'knowledge of representation'. For the assumption is that, by representing knowledge in a system, i.e. by presenting (inputting) the 'object' knowledge in the form of another object—'representation'—to the system, the latter thereby 'gains', 'acquires' or simply 'has' knowledge. Just as the conception of knowledge as representation presupposes that there is a world or something 'out there' to be accurately represented, so too the idea that a system is supposed to contain a representation of knowledge assumes that there is somewhere 'knowledge' to be accurately represented in the memory and/or procedural specifications of the system. Since knowledge is itself representation, representation of knowledge is representation of representation. Thus conceived, the problem becomes a particular case of the general problem of translation from one representational (or semiotic) 'code' into another (recall Locke's definition of 'semiotics' as being, first and foremost, a theory of 'ideas' conceived as signs). Notoriously, many of the discussions within AI (and also within 'cognitive science') of the prob-

* After my oral presentation of the first version of this paper (in December 1986), I became aware of similar—though not identical—attempts to apply the philosophical critique of the concept of representation to AI. An example of this trend is the series of articles by Dreyfus and Dreyfus [15, 16]. Also within AI itself the tendency to look for philosophical partners that depart from the analytic tradition is felt, alongside with a criticism of the assumptions borrowed from traditional epistemology (e.g. [34]).

lem of representation, are indeed explicitly concerned with the design of codes and procedures that ensure the accuracy, reliability, accessibility and usefulness of such translations.* For example, the often raised question of whether, given a visual input, the system should store the knowledge thus 'acquired' in a propositional or in an analog form, is in fact the question of whether the system should map representations in one mode or code onto representations in another, non-isomorphic, code.

Though it is assumed that at some point such a web of representations is somehow anchored in the real world, such an assumption is at most a boundary condition. Within the universe of discourse thus created, all there is are representations and relations between representations, and it is only by reference to the examination of such entities that the 'knowledge' a system is supposed to have can be assessed. Even the correctness of an observational result (itself a representation) is assessed not by reaching out to the world, but through calibration, i.e. by carefully comparing it with the results of different observers, possibly using different observation techniques. That is, it consists ultimately in an *intra*-representational affair. And the question of whether, within the universe of representations, some of them are *intrinsically* better than others, is a notoriously difficult question.† Furthermore, it is well-known that no notation or, for that matter, representation, is self-explanatory or transparent or evident. Its interpretation depends upon 'conventions' or 'rules'. If we assume that such rules also work by virtue of being representations, we end up with an infinite regress.‡ As long as we are thus locked within a representational conception of knowledge, it is illusory to expect that an examination of representations will, in and by itself, provide 'knowledge'—or even a 'gain in knowledge'—of something (say, 'the world') other than of the representations themselves. And this is exactly what 'the representation of knowledge' provides, namely 'knowledge of representations'.

(ii) Since the models or representations are 'computational', the knowledge derived from inspecting them is a knowledge of their 'machinery'. To be sure, the relevant 'machinery' cannot be reduced to hardware alone, but this is immaterial to the identification of this view with the hope of deriving a 'theory of knowledge' from a 'para-mechanical' account of the workings of the mind. AI[B] is thus committing the same kind of confusion between causal explanation and justification as the one attributed by Rorty to traditional epistemology.

(iii) AI[B] also seems to share traditional epistemology's view that there is (or ought to be) a general, 'neutral' set of procedures and/or representations whose inclusion in earlier 'stupid' programs would make them more intelligent. Thus, Boden claims that the deficiencies of Colby's computer simulation of a neurotic 'depend heavily on the lack of a background representation of meaning, or semantics, by reference to which any increased reasoning power that might be supplied would have to function' [5, p. 55] and suggests that the program 'could in principle be radically improved by providing it with ways of representing and using knowledge that are required for intelligent processes in general' [5, p. 33].

(iv) Unlike epistemology, AI[B] does not purport to offer a theory of knowledge. But, as we have seen, at least in one particular case (the knowledge of cognitive processes) it assumes that it is possible to ground knowledge claims on knowledge of representations. Furthermore, in employing the expression 'representation of knowledge' it assumes that somewhere there is a theory of knowledge, on the basis of which the programmer or the system are able to pick out the items to be represented. The methods employed by the system (or the programmer) to identify and 'extract' such knowledge from 'the world', from textbooks, or simply from a data base, are themselves an embodiment of a set of assumptions about what is knowledge and how it is obtained. In all likelihood, these implicit assumptions are nothing but those of the traditional conception of knowledge as an assemblage of representations, and can therefore be criticized on the same grounds.

SECTION 4

The narrower conception of AI (AI[N]) does not, at least ostensively, share the theoretical goals of AI[B] and of epistemology. It thus avoids some of the above criticisms. Given the variety of programs already available, as well as their notorious eclec-

* For an example of this approach, see [4].

† See [22] for an attempt (which I believe fails) to provide strictly internal criteria for choice between competing 'versions' (i.e. representations), which is how he describes our systems of beliefs and our theories—in short all of our 'knowledge'.

‡ Philosophical arguments to this effect have been provided, as is well-known, by Wittgenstein. Some AI researchers (e.g. [35]) are aware of the problem of the non-transparency of notation, and have often called attention to it. Amarel [2] compares the power of several forms of representation (graphical and other) as a means to foster the solution of the cannibals and missionaries problem. His analysis may suggest that some forms of representation are better than others because they reflect fully and naturally all the aspects of the problem. But this is not the case. First, each of the representations discussed by him depend on interpretive conventions just as any of its competitors. Second, the superiority of some forms of representation stems from the fact that they correspond more closely to some other representation (verbal, for instance), taken to be the 'correct' representation of 'the problem'. Boden's [5, pp. 334–340] lengthy and valuable discussion of Amarel's work overlooks these points.

ticism, it is hard to generalize. Yet, in so far as the more 'intelligent' programs make use of the notions of knowledge, representation, and cognates, certain recurrent problems in the field can be traced back to the uncritical acceptance of familiar epistemological dogmas.

Unlike AI[B], AI[N] aims at developing specific tools to perform fairly specific tasks. An example of this approach is the development of 'expert systems' which specialize in the solution of a certain range of problems in a specific field of application. The enthusiasm for these systems derived mainly from the belief that by restricting the field of application and the range of problems it would be possible to overcome the need to supply the system with the enormous amounts of general or background knowledge required for 'general' intelligent behavior. Instead, it seemed possible to restrict oneself to the specific knowledge belonging to the field of application of the system. Likewise, it was assumed that the system will not need general 'thinking power', but only those specific 'inference rules' that proved useful in the intended field of application. It now appears that such hopes were, to say the least, exaggerated. To be sure, some of the expert systems so far developed have become operative. But their level of performance nowhere approaches that of the human experts they were designed to emulate and ultimately replace, and on whose knowledge they were built. At best, they reach the level of modest 'competence'.*

One of the reasons for the relative failure of the expert-systems approach so far has been the tendency of the designers of these 'modest' systems to adopt both the aims and the procedures of AI[B]. The very axiom that such systems must be 'intelligent' and that, for that purpose, in spite of their context-specificity, they should include a general 'inference machine' of some sort, assumes—much like AI[B]—that 'intelligence' has at least some general characteristics, applicable to all contexts, that should be most efficiently described and programmed in a 'neutral', context- and application-independent way. After all, as the programmers of an expert system for the diagnosis of computer failures have recently discovered, the diagnosis of computer malfunctions is just a specific case of 'diagnosis in general', for which there are (or should be) general methods. Clearly, this is a return, albeit unawares, to the grandiose projects, common in the early days of AI, of building a 'general problem solver'. But in so doing, the designers hardly come closer to their purported aim of identifying and imitating the expert's expertise, which presumably does not lie in their ability to employ those reasoning procedures that are common to every domain.

This tendency to rely on general schemata not only cancels out the original emphasis of AI[N] on specificity. It also brings about the adoption of the most problematic assumptions of AI[B]. A peculiar character, from this point of view, is that of the 'knowledge-engineer', a key figure in the design of expert systems. This is someone who is able to 'milk' the human experts in any field so as to obtain in this way the specific knowledge base necessary to fill in the content-empty (and therefore 'neutral') 'knowledge-shell' of an expert system. The knowledge-engineer is not an expert in this or that field of application. He is rather a *knowledge expert.* That is to say, his task is very similar to that of the traditional epistemologist: he knows how to distinguish between what is knowledge and what not, he is supposed to disclose what is the 'method' underlying the experts' achievements in their fields, and he possesses a universal framework that allows for the adequate representation of any kind of specific knowledge. These are precisely those assumptions of traditional epistemology most questioned by current critics. Obviously, the knowledge-engineer, if he is to survive, must be able to cope with the very same criticism aimed at his philosophical counterpart.

But let me discuss and illustrate more closely two interrelated and recurrent problems faced by AI narrowly conceived.

(i) *Dealing with context*

It is widely acknowledged that a 'stupid' program can be made to avoid some of its mistakes by endowing it with some 'understanding' (of its ends, inputs, means, etc.), with the abilities to learn and draw inferences (rather than merely looking up lists of stored knowledge), etc. As noted before, all this requires the ability to view a particular input, result, or procedure within a broader 'context'. That is to say, the more a system operates in a less atomistic and bottom-up way, and becomes more holistically and top-down oriented, the more it behaves 'intelligently'.*

* Dreyfus and Dreyfus [15, 16] distinguish between five levels of performance of a human being (or a system) acquiring specific skills. They claim that existing 'expert systems' do not reach the highest level—that of the *human expert*—but only the third level in their scale. And this, not for technical contingent obstacles, but rather for principled reasons.

* Some authors define not only understanding but also meaning in terms of context-integration (e.g. [19, p. 232n]). For a discussion of the strategies of understanding underlying this and other proposals, see [9]. In spite of its immense importance, the context cannot be the only factor in the process of (linguistic) comprehension. Another factor, not by itself sufficient but certainly necessary, is 'literal meaning' (*pace* all of its recent detractors), which is a function of the semantic and syntactic rules of a language [11].

But 'context' is a rather slippery notion. What is, for instance, the contextual information required in order to understand an utterance of 'Madam, you weigh just over 200 pounds'? In addition to knowing who is the speaker and the addressee, we must also know whether the sentence was uttered in the context of a monthly check-up at the Control Your Weight Association or else in a crowded bus where the lady in question is standing on the speaker's foot. Only in the light of such information can we decide whether the utterance expresses a statement or a request. And if the context is the former one, we will be able to determine whether the statement also expresses a positive or a negative evaluation if we know how much she weighed in the preceding check-up. It is clear, thus, that the 'context' cannot be reduced to a set of standard parameters which can be established in advance for every possible utterance, for in fact any piece of information may become relevant for the proper understanding of an utterance.

In order to reduce this apparently endless proliferation of contextual factors, it is customary to package them in tightly organized structures known variously as 'scripts', 'frames', 'schemata', 'semantic networks', etc. These structures organize pieces of related information and allow for interpreting 'intelligently' input for which they are appropriate. For example, the lifting of her arm by someone sitting in an auditorium is interpreted as a request for permission to ask a question, if the script is 'Lecture', whereas it may be interpreted as an expression of support for a given proposal if the appropriate script is 'Voting'.*

In spite of its usefulness, this way of handling the context poses serious problems. From the last example, we see that a system, except when designed for very narrowly defined purposes, cannot *always* use the same script or schema. If it is not to behave stupidly, it must be able to shift from one schema to another, when required. But how is the system to know when this *is* required? Were it to use a meta-script (or meta-context), the same problem would arise at the higher level. Apparently, scripts and similar 'knowledge packages' are still too atomistic in nature to enable a system to cope in a truly holistic way with an unpredictable multiplicity of situations. Humans—it is often argued—do so by employing not 'meta-schemata' (or meta-rules), but rather by relying upon 'tacit' procedures, such as 'intuitions', non-reducible to 'rules' or similar constructs. But these, it is further argued, are by their very nature *essentially* unarticulated and implicit.* If this is indeed the case, then human understanding (which is the prototype of all understanding) is not amenable *in principle* to suitable computational 'representation' in terms of the procedures currently employed in AI.

(ii) *Imputation*

A related problem is due to the fact that, by letting the system 'understand' too much (say, a given input) in terms of the system's stock of packaged knowledge, the result may be that it will, instead, 'misunderstand', without being able—as humans normally are—to detect and correct such misunderstandings.

This is dramatically illustrated in the case of ascribing propositional attitudes, such as belief, to agents on the basis of actions (e.g. speech acts) performed by the agents. Any rendering of an input sentence like 'John believes that the water is boiling' that uses an analysis of the terms in question (based on the system's semantic or factual knowledge), such as 'John believes that the H_2O is forcibly expelling vapor' risks to wrongly impute to the agent a belief (or knowledge) he does not hold (since, say, he never learned any chemistry).

But it is important to stress that the problem is quite general. First, it does not arise only by virtue of the gap between the technical knowledge available to the system and to John. If the system employs *any* definition of 'water' and 'boiling', there is always the possibility that John himself is not aware of that definition or does not employ it on that occasion. Second, the problem does not arise only in contexts philosophers call 'intensional', illustrated by the use of psychological verbs such as 'believes', 'desires', 'thinks', etc. The understanding of any speech act (or of any action) involves ascribing intentions to the agent, and such an ascription can always fail by imputing to the agent intentions he or she does not actually have.† Third, even for non-intentional events, representing always implies interpreting; in so far as the representation is not a full copy of the represented input, it is based on a selection of those aspects of the input which are considered *relevant by the system's lights*, and hence may be 'incorrect' by some other standard.

* Scripts (and similar constructs) are appropriate for handling certain kinds of contextual factors, but not all of them. For a possible taxonomy of the types of context involved in the understanding of speech acts and texts, see [13]. This may be compared with [7].

* For this claim, see for example Searle's [32] notion of 'background'. The expressions 'intuition' (Bergson), 'tacit knowledge' (Polanyi), and 'tradition' (Popper) refer to modes of knowledge which share this irreducible implicit character.

† Nevertheless, humans usually are able to ascribe the 'correct' intentions in normal verbal or non-verbal interaction. That is to say, they are able, in the case of verbal communication, to reach a reasonable decision as to what is the 'speaker's meaning', even when it does not coincide neither with the 'sentence meaning' nor with the 'utterance meaning'. Pragmatics is the discipline that attempts to explain how this achievement is possible [10].

This is just another way to repeat what has already been argued for, namely that no representation or notation is 'neutral' or 'unbiased', nor, for that matter, self-explanatory or 'transparent'. All representation not only requires, but also *is* interpretation, and therefore involves 'imputation' either to the 'world' or to agents that hold beliefs or other attitudes towards the world of certain 'structures'. This thesis amounts to the acknowledgment that there are no intelligible essences to be 'discovered' out there, but that we (humans, systems) are those who construct intelligibility, according to our changing expectations and purposes. That is, it amounts to recognizing the inevitable subjectivity of our 'representations', i.e. 'interpretations'. Now, the ideal of constructing a system of representation that avoids imputations amounts to the idea—characteristic of traditional epistemology—of providing or finding a 'neutral', mirror-like representation of 'the world' (or part of it). In the light of the criticism to which this view has been subjected, it would perhaps be wise for AI to give up such an ideal, and take advantage, instead, of the pragmatic constraints that guide our 'reasonable' imputations.

It has been suggested that the imputation problem can be solved—at least in this particular case—by the strategy of rendering inputs of the form 'X believes that P' in terms of representations of the form 'X's data base contains X's internal rendering of P' [3]. This strategy in fact says that, since the problem of imputation arises from relying upon explication or analysis of meaning and its use in synthesizing complex representations, one should refrain from synthesizing anything, and thus avoid having to resort to any analysis or explication. The underlying idea is that one's rendering or representation of a situation should be as close as possible to a copy thereof, thus avoiding any 'interpretation' in the representing process. In the case of belief, this is done by assigning to the described situation (in this case a putative belief of a system or agent) only (a) the ability to produce some unspecified 'internal rendering' of an input sentence and (b) the possession of a list (data base) of internal renderings to which the new rendering may or may not belong.

But the system thus rendering such inputs, though designed to avoid illicit imputations, would in fact be imputing to agents a very specific 'psychology'. Two components of this implicit psychology deserve to be mentioned: (a) some of the internal representations or renderings of the system are 'privileged') by virtue of belonging to a 'data base'; (b) the system can have beliefs (or, for that matter, knowledge) even though it is not able to 'understand' (i.e. to analyze or explicate) the meanings of the renderings of such beliefs.

Consider first the second of these components. Unless the 'internal rendering' corresponds to something that might be called an 'understanding' of P, it wouldn't make sense to consider it sufficient for ascribing to the system a belief, however unspecified it may be. Suppose the internal renderings of sentences by a system are merely phonological representations thereof. Suppose, further, that the phonological representation of a sentence Q is in the data base of the system, or else can be derived from representations which are in the data base by means of phonological rules available to the system. Can one say that, in this case, the system 'believes that Q'? Obviously, for the answer to be positive, one should also require that the internal rendering (the phonological representation) should also have some 'semantic import'.* This can be achieved either by directly listing truth conditions for each sentence or else by compositionally deriving such conditions from the truth-conditions of the sentences listed in the data base. The former strategy would severely limit the system's ability to hold an unlimited number of beliefs, and the latter would bring us back to the synthesis of the meanings of complex sentences out of the meanings of their components, i.e. to the very procedure that caused the problem of imputation in the first place. A further problem is due to the general operationalist flavor of the proposal: there is *one* procedure for ascribing beliefs to the system, namely checking the list contained in the data base. But any other such operational definition or disjunction thereof would suffer from the same drawback. A 'truly believing' system may have a non-specifiable number of ways (or operations) through which it can come to hold a belief. The specific way or procedure through which the belief was arrived at does not exhaust its content nor its possible representations for us. In fact, we find here the confusion between causal accounts and justificatory accounts of knowledge again. Beliefs are not such because they are reached by some specified causal process (or by any such process picked out from a complete list of permissible processes). They are what they are because agents that entertain them also entertain (or reject) other beliefs (as well as other propositional attitudes). In short, they are what they are by virtue of being part of a complex network of relations to other similar entities. These relations are of a 'justificatory' or 'opposing/critical' nature. They are not only or mainly causal–internal, but rather social–external. A system has beliefs if it is able to play reasonably well

*Human beings are sometimes required to believe sentences that they are, in principle, unable to understand. A classical example of this situation is that of the so-called 'mysteries of faith' of certain religions [12].

the justification game (so as to let it defend its beliefs in the face of reasonable criticism, and not by virtue of possessing internal representations which it inspects.

As for the first component: the implicit psychology that makes knowledge or belief dependent upon the existence of a 'box' (in our heads or in the system's memory) with the shining label *DATA*, is closely connected with the assumptions of traditional epistemology that have been criticized above. It takes the atoms of knowledge to be representations marked as privileged by some given criterion. The picture of the mind underlying this conception is that of a big container full of representations, some more accurate than others,* the former being kept in a special safe. In epistemology, their privilege derives from their being 'true', i.e. accurate representations of 'reality'. Here, however, their privileged status may derive from some other form of authority: their being those representations that have been fed into the system by the programmer, by the 'knowledge engineer', or by the user. From the point of view of the system, all these are external sources of authority, functioning as *dei ex machina*, not to be contested. In principle, such a system is passive *vis-à-vis* what is 'outside': just like the classical mind, it is a *tabula rasa* which should reflect as accurately as possible what is presented to it; any real activity on its part is likely to distort what it should represent through interpretation or imputing; hence, it is not surprising that its only legitimate activity is the inspection of the representations it contains.†

It seems to me that expert systems and other examples of AI[N] are doomed to face at some point both the problem of the context and the problem of imputing. In so far as in trying to cope with these problems researchers in these fields continue to pay allegiance to the problematic traditional assumptions, I don't see any reason to endorse their optimistic claims about the prospects of their work.

SECTION 5

I will conclude by sketching an alternative conception of knowledge that emphasizes the praxis of justification rather than the notion of representation, and provides a better framework for analyzing 'intelligent' performance.

In Plato we already find the proposal to distinguish between mere belief and knowledge (*episteme*) through the addition of two conditions: truth and justification. According to this proposal, somebody knows that P if, and only if, she believes that P, P is true, and her belief that P is justified. The value of this proposal depends, of course, on the interpretation of 'justified'. An excessively broad interpretation may qualify as knowledge beliefs that are perhaps true, but dubiously grounded (John *knows* that there is life after death because it says so in the scripture; Mary *knows* Michael is the murderer because she saw blood stains in his coat; system X *knows* that P because P is listed in its data base). An excessively narrow interpretation, on the other hand, requiring, say, no less than certainty, would disqualify as knowledge many beliefs which it is unreasonable to doubt (John *doesn't know* that he has a hand because his visual and tactile sensations, upon which his belief that he has a hand is based, may be the result of hallucination; the judge *doesn't know* that Michael is the murderer because it is always possible that all the evidence against him has been 'planted'; a system X *doesn't know* that P because, in spite of all the safety devices, one cannot guarantee that the code representing 'P' has not been inadvertently changed into the code representing 'Q' in the course of a power failure).

Whether adopting more or less stringent requirements, traditional epistemology attempted to explain the concept of justification in one of two ways: logically or psychologically. All the traditional theories of knowledge share the assumption that there are pieces of knowledge which do not require justification at all because they are 'evident', be it because they stem from immediate sensation (Locke), or because they are perceived as such by direct intellectual intuition (Descartes), or else because they constitute a necessary condition of possibility for our experience of the world and intellectual activity (Kant). They also share the assumption that the qualification of any other belief as knowledge depends upon the way in which it is 'derived' from the evident beliefs. It is at this point that the traditional theories diverge in their construal of justification. For some of them, justification is accounted for in terms of generation through 'acceptable' psychological processes (habit, conditioning, generalization, production of complex ideas out of simpler ones, and so on). For

* For a description of the network of frozen metaphors that express this picture of the mind, see [27].

† It is worthwhile noting the analogy between this picture and a certain conception of education dubbed by Paulo Freire 'the banking concept of education':

> Education . . . becomes an act of depositing, in which the students are the depositories and the teacher is the depositor. Instead of communicating, the teacher issues communiqués and 'makes deposits', which the students patiently receive, memorize, and repeat. This is the 'banking' concept of education, in which the scope of action allowed to the students extends only as far as receiving, filing, and storing the deposits. [20, p. 45]

I would conjecture that the reason why many of the educational applications of AI do not succeed in fulfilling their promise of developing creativity in the students is precisely the fact that they rely upon the 'banking concept of knowledge' I have been trying to describe.

others, justification is rather accounted for in terms of the possibility of grounding non-evident beliefs upon evident ones by means of logically valid connections (proof, verification, consistency, and so on). But in both cases justification is conceived as reducible to a definite set of 'legitimate' moves (be they logical or psychological–causal) which lead from that which is already justified (or not in need of justification) to that which is to be justified. And in both cases the issue of the form of representation (or, to use AI's language, the 'format') is crucial, since the moves in question must be *general* in nature and relatively few in number, which means that they must be defined *formally*. That is to say, their legitimacy must be a function of the form of the representation, and not of its content.* There is no need to stress that both AI[B] and AI[N] fully endorse the conception of knowledge just described.

But there are powerful arguments against the reconstruction of the justification relation either as a purely logical or as a purely psychological relation. The anti-psychologistic arguments of thinkers such as Husserl, Frege, and the logical positivists have shown that the psychological origin of our beliefs—no matter how 'acceptable' the psychological processes that lead to them turn out to be—offers no guarantee whatsoever of their reliability. Cognitive psychology supports these anti-psychologistic claims, by providing many experimental results that convincingly show that our 'natural thinking' systematically violates the simplest rules of logic.† On the other hand, the failure of the most elaborate attempts to provide a decent logical formulation of the positivistic 'verification criterion', as well as the well-known counter-examples against the analysis of knowledge as justified true belief, have shown that the justification cannot be conceived as a purely logical relation.‡

Let us then leave aside the logical and psychological paradigms, and look for another way of construing justification, a *social* one. It is convenient to recall that the praxis where the concept of justification originates is judicial (see Romano [28]). In terms of this praxis, to justify or to prove guilt is to remove all *reasonable* doubt. By 'reasonable doubt' it is understood any *bona fide* doubt that the judge or the jury are likely to raise in the context of the trial. We are talking, therefore, about the ability to *convince* certain persons in a certain context, against the background of certain beliefs which are not in doubt (in the context), and which serve as the basis to assess the reasonableness of any given doubt. Thus, for example, if the context is an inquisitorial trial in the 16th century, doubts concerning the existence of the devil and its ability to take possession of human souls are not 'reasonable'. Consequently, they are not acceptable as justifications on the part of the defence. Similarly, regarding current claims of knowledge outside of the courts: general skepticism, of a philosophical nature, about the reliability of our senses as a source of information about the world is not a reasonable justification, under normal conditions, against the claim that I know that I see in front of me a computer, that my fingers pressing the keys in the keyboard operate it, and so on. The proviso 'under normal conditions' is important. It drives us back to the 'context'. For there can be contexts where there are good reasons to doubt the testimony of our senses and of our common knowledge (the computer might be secretly operated by an occasionalist evil genius or 'virus' that wants me to believe, falsely, that I am in charge). Reasonableness is not defined in a universal way, for all possible contexts, but it is an essentially context-dependent concept.

The conception of knowledge that emerges from the above account is, thus, a relativized one. Not in the sense that would require us to reply to all questions of the type 'Do you know that P?' by an overly cautious 'It depends'. In many circumstances we are able to and should reply, without special difficulty or hesitation, 'Yes' or 'No'. We will answer 'Yes' when we possess justifications against *reasonable* alternatives that have been raised or that are likely to be raised. Knowledge is 'relative' because in another context, the set of reasonable alternatives may be different, so that the justifications we possess may not meet them. Hence, it would be better to call this conception of knowledge a 'pragmatic' one, as suggested by Dretske [14]. This label has a further advantage: it calls attention to the fact that the justification upon which knowledge claims depend is, first and foremost, a *communicative* (hence, social) activity, rather than a purely *mental* (hence, private) one. A system which is able to justify its beliefs is a system which is able to *defend* them against objections. In order to do so,

* On the need for such a 'formality condition' within the framework of a representational/computational model of cognitive processes, see [18].

† See, for instance, [17] and [24].

‡ On the difficulties in formulating the verification criterion, see [8]. Gettier's [21] counter-examples are of the form: David sees an animal on the hill, which he identifies easily as a sheep; the statement P ('There is a sheep on the hill') is true, because indeed there is a sheep on the hill; David believes that P; and David has a justification for his belief (he *sees* there a sheep); all the conditions of knowledge being thus fulfilled, one may say, on the classical account, that David *knows* that P; but it turns out that the animal he sees is not a sheep, but a goat; the sheep that ensures the truth of P is another animal, grazing behind a tree, and thus not seen by David; in such a situation, to say that David *knows* that P, contradicts our basic intuitions, in spite of the logically flawless argument supporting that claim. Many modifications of the classical account have been suggested in order to cope with examples of this kind (see, for instance, [6]), but most of them are flawed by the reliance on pure logical deduction as a means of construing the justification relation.

it must *understand* the objectors' claims, and to formulate its own claims and reasons in such a way that the objector will understand them. It is immaterial whether the system actually performs all these steps. It can just 'plan' them in its 'head'. Still, it is constrained by the pragmatic principles that govern the intelligibility of all forms of communication.

This is not the place to present in detail what we can learn from pragmatics about the nature of such constraints.* Let me recall the main points: understanding cannot rest solely on the 'semantic' decoding of the signs (linguistic or other) that the system receives as input or otherwise manipulates; it is necessary—even in the simplest and most 'transparent' or 'literal' speech acts—to take into account the context (and not only the 'immediate' context) as a factor which is capable of changing completely the initial semantic interpretation (i.e. the conventional meaning) of those signs. As a consequence, understanding—and with it, justification too—cannot be reduced to the application of semantic or logical algorithms alone, no matter how sophisticated they may be. Instead, it must be conceived as a heuristic process whereby interpretive hypotheses are raised, followed by reasonable objections to them in the light of contextual information, and concluding with the choice of the most reasonable interpretive hypothesis in the given context. In this process, the role of the semantic interpretation is, to be sure, important: it functions as a first hypothesis, whose adequacy or not to the context leads either to the immediate conclusion of the process or to the search for other hypotheses. But even if the process is terminated by accepting the semantic interpretation, i.e. even if one concludes that the speaker has conveyed precisely what his words convey, such a decision is itself context-dependent and hence heuristically achieved. For what it amounts to is simply the acknowledgment (by the interpreter) that no reason was found (in the light of his contextual knowledge) to doubt the 'literalness' of the speaker's speech act.

Just as it is impossible to get rid of the context, so too it is impossible to avoid imputation. Every understanding involves imputing an interpretation to a given speech act (or to any other semiotic act). And since every such interpretation is dependent upon a context whose boundaries are never fixed once and for all (nor beforehand), imputation is always fallible. In addition to the failed attempt already discussed to avoid any imputation, two ways are open to those who want to prevent 'imputation mistakes' once and for all: (a) to rely only upon the conventional/algorithmic interpretation; and (b) to reach the contextual interpretation via a *complete* search of all contextual factors. Both ways are self-defeating. The former, because it means giving up real understanding, which is always contextual, thus leading back to 'stupid' or 'blind' systems. The latter, because it implies the system's paralization, given the potential endlessness of the context. We must therefore accept the inevitable fact that all imputation is necessarily tentative and fallible. All that it is reasonable to demand is that it be . . . 'reasonable'.

Here we are, again, facing this awkward and vague concept of reasonableness. This notion, just as its inseparable fellow, the notion of relevance, notoriously escapes all attempts to render it more precise and explicit. The use of non-monotonic logics, for example, which allow for the derivation of refutable conclusions, no doubt captures one of the important characteristics of reasonableness. But even amongst refutable inference rules, contradictions that can only be resolved by further appeals to the context can arise. For instance, the temporal projection rule, which says that if X has a certain property at t_0, then it has the same property at t_1, may generate contradictory predictions: (a) if gun X is loaded at t_0 then (defeasibly) it is loaded in t_1; (b) if person A is alive in t_1 then (defeasibly) she is alive in t_2; (c) gun X is fired with precision towards A at t_1. Whereas from (a) and (b) it is reasonable to conclude that A is alive at t_2, from (a) and (c) it is equally reasonable to conclude that A is dead at t_2. According to Hanks and McDermott [23], the latter conclusion is more reasonable than the former. They account for this alleged fact in terms of a meta-rule which orders temporal projections in terms of the chronological order of the events in question. But this meta-rule is itself problematic, as has been shown by Loui [25]. One of the problems is that each of these inference rules and the meta-rule is dependent upon the prior delimitation of a range of 'reasonable' possibilities which are taken into account:* if one

* For a detailed account, see [10].

* It seems to me that Loui in fact accuses Hanks and McDermott of committing 'Locke's mistake', which was described above. The conceptual mistake he ascribes to them is this:

> Facts are facts, and they are timeless. Beliefs, on the other hand, have the kind of inertia that Hanks and McDermott envision. We tend to hold beliefs until we are forced to relinquish them. . . . We should be legislating rational beliefs about temporal relations among properties. That's not the same thing as reproducing the temporal properties of an agent's belief-forming processes. We can state the mistake as a transposition: there is a difference between the temporal evolution of beliefs, and beliefs about temporal evolution. [25, p. 296–297]

The criticism against Hanks and McDermott's excessive psychologism seems justified. Yet Loui's own position, according to which one has to replace the rules proposed by them by rules which are better because they *represent* more accurately not the psychological processes of belief formation but the facts about temporal succession, shows that he is still playing the representational game of classical epistemology.

considers seriously the possibility that between t_0 and t_1 the gun was unloaded, then the meta-rule cannot any longer provide a justification for the superiority of the conclusion that A is not alive at t_2. When the range of possibilities is not defined beforehand, i.e. when there is no precise criterion of relevance that 'closes' the domain of alternatives, then whoever is to draw a reasonable inference must not only employ the given rules and meta-rules, but also decide whether it is reasonable to employ them at all.

Somehow we humans succeed in guiding by reason our actions in an environment which is not only non-deterministic, but also requires the use of criteria of reasonableness and relevance which change very often and even contradict each other. 'Intelligent' or 'rational' behavior in such an environment means the ability to choose in every context the appropriate criteria. Otherwise, the behavior will be 'blind' or 'dogmatic', and it will in all likelihood be self-defeating in the long run. There is no doubt that scripts, frames, and similar structures are locally useful. But their use should be subordinated to the ability to re-assess continuously the situation as a whole, a reassessment that may eventually show that the frame or script previously chosen turns out to be inadequate, and must be replaced by another. When Abelson's 'Ideolog' concludes that it is highly plausible that Fidel Castro will throw eggs at Taiwan, based on his knowledge that Fidel Castro is a communist, Taiwan's regime is anti-communist, and radical students in Venezuela have thrown eggs at Nixon the anti-communist, the program's inference fails to provide a 'reasonable' justification for its conclusion. The reason for its failure lies not only in its lack of common sense background knowledge (such as the geographical knowledge about the distance between Taiwan and Cuba, or the anatomical/physical knowledge about the range of manual egg throwing), as has been argued by the program's designer [1, p. 274]. The reason lies also in the fact that the program applies blindly—that is, not 'intelligently'—a single script for the prediction and explanation of *all* activities of those agents which fall under a certain ideological classification.* The failure of the program to provide a model that replicates the behavior even of people that are indeed prey to some ideology is a consequence of the fact that it imputes to *them* the very same blindness that characterizes it.

If indeed, as I have tried to show, the ability to adapt pragmatically to changes in the context is one of the main characteristics of intelligence and rationality, and if 'knowledge' is one of the means for intelligent behavior, then the crucial question is not that of 'representation of knowledge', nor that of 'providing (more) knowledge' to a system. Rather, it is the question of the possibility of designing systems that are not enslaved by something labelled 'knowledge', i.e. systems which are able to reject justifications that do not seem reasonable to them, and to select pragmatically even the criteria themselves of what is to be considered, in each context, as 'reasonable' and 'relevant'. Researchers in AI should direct their attention to the question of whether it is possible to develop systems which are not subordinated to the knowledge and to the rules and criteria which are supplied to them *ex machina*, and if so, how. And they should not forget that this pragmatic aspect of knowledge derives from the public/social character of all justification. The real Turing test for an 'intelligent' system is the ability to provide convincing justifications for its responses and performances. Nowadays, the potential candidates for being convinced are the human beings that interact with the systems. But tomorrow, with the development of techniques of parallel processing and inter-computer communication, it is perfectly possible that the systems' persuasive powers will be addressed mainly to their fellows in a 'community of computerized systems'. Maybe in this new society the standards of relevance and reasonableness will be quite different from ours, but it doesn't seem likely that their essential dependence upon contextually changing circumstances will thereby disappear.

* Even when ideology is conceived in very broad terms [26, pp. 49–53]— in the sense that it is supposed to explain all the actions of the members of a given social class—it is preposterous to assume that it is in terms of ideology that one has to understand and predict *all* their actions, with no exception: can one explain or predict the activities connected with bicycle riding in terms of class ideology?

REFERENCES

1. R. P. Abelson, Concepts for representing mundane reality in plans. In D. G. Bobrow and A. Collins (eds), *Representation and Understanding: Studies in Cognitive Science*, pp. 273–309. Academic Press, New York (1975).
2. S. Amarel, On representations of problems of reasoning about actions. In D. Michie (ed.), *Machine Intelligence* 3, pp. 131–172. Edinburgh University Press, Edinburgh, U.K. (1968).
3. J. A. Barnden, Imputations and explications: representational problems in treatments of propositional attitudes. *Cognitive Sci.* **10** (1986), 319–364.
4. D. G. Bobrow, Dimensions of representation. In D. G. Bobrow and A. Collins (eds), *Representation and Understanding: Studies in Cognitive Science*, pp. 1–34. Academic Press, New York (1975).
5. M. Boden, *Artificial Intelligence and Natural Man*. Basic Books, New York (1977).
6. R. M. Chisholm, *Theory of Knowledge*, 2nd edn. Prentice-Hall, Englewood Cliffs, N.J. (1977).
7. H. H. Clark and T. B. Carlson, Context for comprehension. In J. Long and A. Baddeley (eds), *Attention and Performance IX*, pp. 313–330. Lawrence Erlbaum, Hillsdale, NJ (1981).

8. M. Dascal, Empirical significance and relevance. *Philosophia* **1** (1971), 81–106.
9. M. Dascal, Strategies of understanding. In H. Parret and J. Bouveresse (eds), *Meaning and Understanding*, pp. 327–352. De Gruyter, Berlin (1981).
10. M. Dascal, *Pragmatics and the Philosophy of Mind*, Vol. 1. John Benjamins, Amsterdam (1983).
11. M. Dascal, Defending literal meaning. *Cognitive Sci.* **11** (1987), 259–281.
12. M. Dascal, Reason and the mysteries of faith: Leibniz on the meaning of religious discourse. In M. Dascal, *Leibniz. Language, Signs, and Thought*, pp. 93–124. John Benjamins, Amsterdam (1987).
13. M. Dascal and E. Weizman, Contextual exploitation of interpretation clues in text understanding: an integrated model. In J. Verschueren and M. Bertucelli-Papi (eds), *The Pragmatic Perspective*, pp. 31–46. John Benjamins, Amsterdam (1987).
14. F. I. Dretske, The pragmatic dimension of knowledge. *Phil. Stud.* **40** (1981), 363–378.
15. H. L. Dreyfus and S. E. Dreyfus, Competent systems: the only future for inference-making computers. *Future Generation Comput. Syst.* **2** (1986), 233–243.
16. H. L. Dreyfus and S. E. Dreyfus, Why expert systems do not exhibit expertise. *IEEE Expert* **1** (1986), 86–90.
17. J. St. B. T. Evans, *The Psychology of Deductive Reasoning*. Routledge and Kegan Paul, London (1982).
18. J. Fodor, Methodological solipsism considered as a research strategy in cognitive psychology. In J. Fodor, *Representations: Philosophical Essays on the Foundations of Cognitive Psychology*, pp. 225–253. The MIT Press, Cambridge, MA (1981).
19. J. J. Franks, Towards understanding understanding. In W. B. Weimer and D. S. Palermo (eds), *Cognition and the Symbolic Processes*, pp. 231–262. Wiley, New York (1974).
20. P. Freire, *Pedagogy of the Oppressed*. Penguin, Harmondsworth, U.K. (1971).
21. E. L. Gettier, Is justified true belief knowledge? *Analysis* **25** (1963), 121–123.
22. N. Goodman, *Ways of Worldmaking*. Hacking, Indianapolis, IN (1978).
23. S. Hanks and D. McDermott, Default reasoning, nonmonotonic logics, and the frame problem. *Proceedings of the American Association for Artificial Intelligence*. Philadelphia, PA (1986).
24. D. Kahneman, P. Slovic and A. Tversky (eds), *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, U.K. (1982).
25. R. P. Loui, Response to Hanks and McDermott: Temporal evolution of beliefs and beliefs about temporal evolution. *Cognitive Sci.* **11** (1987), 283–297.
26. K. Mannheim, *Ideology and Utopia*. Routledge and Kegan Paul, London (1936).
27. M. Reddy, The conduit metaphor—a case of frame conflict in our language about language. In A. Ortony (ed.), *Metaphor and Thought*. Cambridge University Press, Cambridge, U.K. (1979).
28. C. Romano, The illegality of philosophy. In A. Cohen and M. Dascal (eds), *The Institution of Philosophy: A Discipline in Crisis?* (in press).
29. R. Rorty, *Philosophy and the Mirror of Nature*. Princeton University Press, Princeton, NJ (1979).
30. R. Rorty, Pragmatism and philosophy. In R. Rorty, *Consequences of Pragmatism*, pp. xii–xlvii. University of Minnesota Press, Minneapolis, MN (1982).
31. J. R. Searle, Minds, brains and programs. *Behav. Brain Sci.* **3** (1980), 417–424.
32. J. R. Searle, The background of meaning. In J. R. Searle, F. Kiefer and M. Bierwisch (eds), *Speech Act Theory and Pragmatics*, pp. 221–232. Reidel, Dordrecht (1980).
33. C. Taylor, Overcoming epistemology. In K. Baynes, J. Bohman, and T. McCarthy (eds), *After Philosophy: End or Transformation?*, pp. 464–488. The MIT Press, Cambridge, MA (1986).
34. T. Winograd and F. Flores, *Understanding Computers and Cognition: A New Foundation for Design*. Addison-Wesley, Reading, MA (1987).
35. W. A. Woods, What's in a link: foundations for semantic networks. In D. G. Bobrow and A. Collins (eds), *Representation and Understanding: Studies in Cognitive Science*, pp. 35–82. Academic Press, New York (1975).