

**AFTER LEIBNIZ... :**



**Discussions on  
Philosophy and Artificial Intelligence**

STANFORD ARTIFICIAL INTELLIGENCE LABORATORY  
MEMO AIM-229

MARCH 1974

COMPUTER SCIENCE DEPARTMENT  
REPORT NO. STAN-CS-74-411

**AFTER LEIBNIZ... :  
Discussions on Philosophy and Artificial Intelligence<sup>1</sup>**

*by*

*D. Bruce Anderson, Thomas O. Binjord,  
Arthur J. Thomas, Richard W. Weyhrauch,  
, Yorick A. Wilks*

Abstract:

**This** is an edited transcript of informal conversations which we have had over recent **months, in** which we looked at some of the issues which seem to arise when **artificial intelligence and** philosophy **meet**. Our aim was to see what might be some of the fundamental principles of **attempts to build** intelligent machines. The major topics covered are the relationship of **AI and** philosophy **and** what **help** they might be to each other: the **mechanisms of natural inference and** deduction; the **question of what kind** of theory of **meaning** would be **involved** in a successful **natural language understanding** program, **and** the nature of models **in** AI research.

<sup>1</sup> We are very grateful to **John McCarthy** for his helpful comments at various stages **during** our discussion.

*The writing of this paper, and some of the research described herein, was supported by the Advanced Research Projects Agency of the Office of the Secretary of Defense under Contract No. DAHC 15-73-C-0435 and National Institutes of Health Contract NIH-MH 06645-12.*

*The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the U.S. Government.*

*Reproduced in the VSA. Available from the National Technical Information Service, Springfield, Virginia 22151.*

## BR UCE

Well, here are some **questions** we might start off with, though they are so vague that we **can** argue over the meaning of the questions, yet alone their answers. What is the relation **between AI and** the traditional studies of **intelligence** viz. philosophy, psychology **and** linguistics? **In** the following senses

1 Would a knowledge of those traditional subjects help me to make an **intelligent program** directly? Can we firm **up** philosophical (or **even** psychological) theories enough to make concrete statements about robots with them?

2 Would it help me indirectly, **for** example in the **sense** that modal logic isn't directly useful since its formalisms are weak, but the logicians' examples are **illuminating** and raise interesting issues?

3 Is it a good heuristic to ignore all such **subjects**, because it would take so **long** to sift through **them** to find something of value **that** in that time you could have discovered it **from** 'first principles? As by analogy, there is not much point in **ploughing** through **Roman** arithmetic if you have just invented zero **and** the **arabic** notation!

4 Now the other way round. Would a working robot, with natural **language input** etc, have **any** effect **on practitioners** of the **traditional disciplines**? I suspect that philosophers would be unaffected, psychologists helped quite a bit, and linguists mostly wiped out.

## RICHARD

Here are **my** answers, **following** the same numbering: 1 No! ! I don't think **50**. There is **even less evidence** to believe that these disciplines have anything **computationally** significant to say **even** granting that they have **clear** "ideas". 2 Yes. At least in some cases they can tell **you** what **not** to do. 3 **In** general I think that it is important to think through **unsolved problems** (or those with disputed "solutions") without being greatly **influenced** by people's previous attempts at the problem. **Studying** another person's blind alleys before you have a collection of your own is probably a waste of energy. Of course that doesn't mean you should work in ignorance of other peoples work, just that if it has not obviously **succeeded** then you should be skeptical of it. **Good** ideas eventually shine through. **In** addition, most work in those traditional areas was before **any** knowledge of any complex but "**well-understood**" and **manipulable** objects like computers. Thus our whole **experience** is **different** from theirs.

## YORICK

I disagree with your remark about "clear ideas", Richard. It seems to me that in many **traditional** disciplines there were people making **important** use of **notions** like "clear and definite procedure" long before the first computers: the behavioral psychologists; Vienna Circle empiricist philosophers etc. I completely disagree with Bruce above about the **relations between** "tough, good little AI", and "vague sloppy old philosophy". Bruce says that maybe philosophy could be suggestive if **only** it could be "firmed up". This is **all** topsy-t **urvy**: it's philosophy that is precise **and** AI that **needs firming** up, as he puts it. Bruce's criteria of firmness are all wrong: **writing** programs is indeed firm, but can be firm

and totally pointless. just like copying out the **phone** directory by hand or **something** equally definitive. Philosophical analysis can become trivial, I'd be the first to agree, but never quite as trivial as mindless **programming**.

Ezra Pound used to say that “after **Leibniz**, a philosopher was a guy who was too damned lazy to work in a laboratory”, and there’s **something** iii that. Often, philosophical analysis done **in isolation** fails to get **any** insights, **insights** that **might** well come from **trying** to build a system. But analysis has all kinds of goodies to offer AI, At the simplest level it could do **something** to **inhibit** AI usages like “**epistemology**” when “logic” is **meant** (usually “**intensional** logic”). The same goes for the AI use of “**Theory of Knowledge**”. If **anyone** says, like **Humpty Dumpty**, that we **can mean anything** we like by words, then all I **can** say is, why pick those philosophically loaded words to **tinker** with in the first place? These are **not** at all **trivial** errors. They **clutter** Up thought.

ARTHUR

I **concur** with what you’ve just said, Yorick, but I’d like to press the **point** further and say that AI people may **simply not** understand that a great **many** of **the** virtues that they see **emanating from** the ‘**computational** metaphor’ **are** already **well-known** to philosophers. Your **example** of ‘clear and definite procedures’ **indeed** goes back to the **notion** of algorithm which obviously was understood by **ancients** like Euclid **and** Pythagoras. Pattern-matching, that current darling of AI, is a **notion** that **Wittgenstein** certainly **knew** about.

A lot of the **confusion** may be caused by a failure to realise that philosophy has evolved a very **definite technical** vocabulary, **and** that, as you said, you just can’t go **around taking** a philosophical **argument**, making the words mean what you **think** they **should** mean, **and** then pronouncing the whole **thing** to be sloppy, **irrelevant** or just **plain** boring!

BRUCE

**Can** you be a little more **definite** about **Wittgenstein and pattern** matching?

ARTHUR

What I **mean** by that is that **Wittgenstein’s** major **concern** in his early work was the **question** of how the logical structure of **propositions** reflects, or more precisely, is a ‘picture’ of the world it represents. His **conclusion**, roughly speaking, was that the **meaning** of a proposition was embodied by the actual **disposition** of variable **and** predicate **names** within it, **and** that **bindings** to the variables must obey category rules which **correspond** to the **categorial** structure of the world. According to this **interpretation**, the reason, for **instance** why the **sentence** ‘Aristotle is stupidity’ is meaningless is that there is **no** legitimate logical form for the corresponding fact.

BRUCE

**Oh**, at first I thought your meant **Wittgenstein knews** all kinds of things about **pattern-matching** larrguages (or sublanguages), **concerning** good **and** bad features, **implementation** problems etc. Now I see **you mean** he had a theory of the world which **depended on some notion** of matching between structures, and that you **think** this could be **implemented** in

terms of pattern-matching as we know it. I think that's a bit different, though still very interesting.

How should we go **on** from here? Our further **discussion** could have two aspects, The first is **technical, and is concerned** with **making** robots. It is about questions such as "how much does a robot have to **know** about the world vs about language?", **and** I think that we are all in agreement about this. There are some misunderstandings, but these are all relatively unimportant. The second component is about things like "what is meaning?". But if we are agreed **on** the first **component, then the second** is really rather **pointless** -- I really don't care which part of the robot's hardware/software is the embodiment of the **meaning** of the word "like", for example. If the robot **can** talk about liking, and reason about it, and seems to "**understand**" the word **in the sense** that someone else does, then I **am** happy. There would **no** doubt be **700** Ph. D. theses **on** whether the **thing** was dictionary-oriented or **not, and** whether it was **Popperian** or **Carnapian** and so on, but these disputes would be restricted to crabby old philosophers and their pupils.

So why **should** we talk about philosophy **and** psychology at all? The reason is that although we are agreed about the types of structures we **need** for our robot, they are nowhere **near** specific/well-defined enough to be implemented. Part of the steam in our disputes is generated by Yorick's **thinking** that the models we are proposing are well-defined but that he does not understand them. **But** really the requirements we have put forward are incredibly vague. We look to philosophy to find *more detailed* ideas about the kinds of structures that could be in our robot. And of course we are **disappointed** when we find lots of philosophers are **even** vaguer **than** we are!

TOM

I disagree with two **things** that Bruce has said we all agree about! First, we are **not** at all **in** agreement about **how** much a robot has to know about the world and about language. Yorick and I, for **example**, are **in** strong disagreement about that, though I think that part of our disagreement comes **from** Yorick **thinking** about **translation** while I **am concerned** with the whole **range** of intelligent behavior.

Nor do we have any **general agreement** about what the structures are. We might agree **on** what behavior is desired. We don't have **agreement on** the design criteria to achieve that behavior, **much** less **on** the specific structures. The **discipline** of AI does have some specific structures but riot enough. arid very few have **long** term value,

YORICK

Well, Tom, I'm **not** sure I want to admit that our disagreement springs from **my** petty **concerns** versus your large **and general interests in intelligent** behavior. The other **person's** **concerns** always look a bit **limited**. It's quite true that I'm more **interested** in language structure (not **in translation per se**) and **how** we understand it, and that **I** assume that the **solution** of those problems will have impact over the whole **range** of human **intelligent** behaviors. Conversely, you're **more** iuterested **in vision and** expect similar general advantages **from any** advance **on the vision** front. **I think** that's a fairer statement of the nature of our disagreement.

Bruce, why are you looking for the philosophers to provide you with programs at all, why should they, what makes you think its their job? Though, I must admit that, having said that, it's clear that the philosophers that AI people tend to be aware of also tend to be those like Turing, Davidson, Grice, Montague and Searle, who have all, in their different ways, provided something like protocols for program. But that doesn't mean the others should, or that those listed are the best or most central philosophers of their generations.

BRUCE

We are not necessarily asking the philosophers to write our programs for us, but to give us constraints on them/ideas about them which are powerful enough to be of (in the best of all possible worlds) direct help. The disappointment comes not because philosophy hasn't solved the "problem" (whatever it is) but because the relevant discussions have not reached a concrete enough level. For example a robot-builder will have to take some position on intensional objects, but the possible positions are hard to find. And I agree strongly that the people who philosophers might point you at if you complained about a lack of concrete <sup>ie5.5</sup> viz. Davidson, Montague etc are even worse as they really have nothing much to say because their formalisms are so puny compared with computational ones.

Philosophy is too concerned with how things *might* be, with reducing possibilities from the top down: but at some stage it is worth diving in and testing out a few ideas about how things are!

ARTHUR

Maybe I'm just not understanding you, Bruce, but I don't see at all how you can say that, for instance, Montague's formalism is 'puny compared to computational ones': where exactly is this powerful computational formalism that you have in mind? After all, you just said that the problem in AI is that the proposed models are not as well-defined as Yorick thinks they are, and that you, in your delightful naivete, have looked to philosophers for more detailed ideas. So how, then, can you turn around and attack a formalism, (like Montague's, however inadequate it might be), after you've admitted that you have nothing to put in its place?

Of course, philosophy is indeed concerned with how things *might* be: but that is precisely its strength! It prevents you from diving in and hitting your head on an unexpected bottom. Drowning is a thing that AI is always in imminent danger of! Part of AI's trouble, as I see it, is that people just don't look before leaping.

BRUCE

But does philosophy tell us that certain approaches to AI will fail? For example, what would the Austin-Strawson-Grice approach to a natural-language understanding program be? Note that (like everyone else interested in philosophy) Yorick and Arthur are pretty schizophrenic in that their philosophy does not enter into their programs, and their writings are clearly split between the practical (cough) and the philosophical. This split is a bad sign -- is there any mutual feedback between philosopher-Wilks and roboticist-Wilks?

## YOR ICK

Well, a Popper-robot **would** be quite different from a robot with a standard philosophy of science: it would go round busying itself **trying** to disprove its general **beliefs** all the time!

## BR CJCE

I still think that getting **into** philosophy, **in** classical philosophy's terms, is probably a waste of **time** (for AI people at least) because the aims and ways of thinking are so different **from** ours, and **digging in** in our **own** terms is so difficult that it has **often been** said (e. g. above) that it is quicker to rediscover any of **anything you** need rather than dig **through** that subject trying to decide what is useful. This is also said of psychology and linguistics, but not of **mathematics** for some **reason** -- probably because so many AI people have **mathematical** backgrounds!

So this is what we ought to be doing, though like the robot's insides it is not specified **enough** to let us just go and do it. **But please let us have less** Of the "**meaning of meaning**" discussion **sort of thing in AI**. Less of the "do read **Strawson on Individuals** it will blow your **mind and** alter your **programming** style immeasurably" stuff too - the latter really **means** "I think **Strawson on Individuals** has **something** to say to AI people, but I'm **damned** if I **know** what it is - maybe you will find out for **me**".

## ARTHUR

I **really can't** accept that I **am**, or **anyone** else **is**, quite as **schizophrenic** as Bruce wants to make out. I **can think of several** occasions **in** the past history of AI where a little critical **forethought in** the philosophical **manner** would have saved people **from** lurching **up blind** alleys. Firstly, for several years (roughly from 1960-68) **most** people who might be regarded as **pioneers of AI** thought that the first-order predicate calculus would be an adequate **vehicle** for representing knowledge for a robot. The result was an enormous effort **in automatic theorem proving** (I'm **not**, of course, saying **that** theorem proving is **not** an interesting **technical** subject). This effort was, from the point of view of people interested **in representation**, of limited usefulness. Now a philosopher could have told you this **beforehand**, simply because he already **understood** the technical limitations **on** what the **language** could possibly express (a standard philosophical **concern, mark you!**) He might **even have been** able to give suggestions for better languages, e.g. **tense logics** for **dealing with** time, change and causality, and **epistemic logics** for **handling** knowledge.

The **second** case is **in linguistics**: a **Chomskian** (and **Chomsky did** think of himself as a **philosopher**) would tell you that **you** needed **transformational** grammars and only a very **little semantics** (probably of the Fodor-Katz flavour) for a **language-understanding** program: and an **ordinary-language philosopher** would (crudely) have said that '**meaning is use**' (**notice the thoroughly proceduralist dictum** - - it could have **been** said by Hewitt!). He would say **that** **what you should do** in the first instance is collect **examples** of how words are used in various **contexts**. A philosopher of the Montague-Hintikka school (**an entity** which I'll **create for** purposes of **argument** -- Montague and Hintikka in fact have **distinctly** divergent views of this subject, but their approach is similar) would argue, as I've tried to, that **what you really need** is a theory of how words refer to objects. I'm not sure, for example whether **Winograd can** be put **into my camp** or with the proceduralists: he seems to have elements of

both. I **don't know** whether he **even** consciously thought about **which position** he was **taking**. I do believe that his work is a good example of **how** a philosophical **position**, **and** one which was **well-known in all its important facets** **several decades** ago, **can be embodied in** 3 powerful program. One **isn't denying** at all the value of **being** able to write programs to test out ideas, but **one** does **want to question** the idea that ideas can evolve by **themselves, ambulando, during the writing** of program. This last **notion** is surely contrary to the canons of programming itself.

BRUCE

It **sounds** to me as though you are **trying** to claim that "critical forethought" is a philosophers' monopoly!

ARTHUR

No, of course I'm **not** trying to claim **anything** like that! All I'm **trying** to say is that **philosophy**, as a **discipline**, is **in** some large part devoted to **trying to analyse** the *a priori* **limits on** our knowledge.

BRUCE

Sorry, it was just **my** way of saying that the **results aren't** obviously more **than** those you **could** get by **using** critical forethought, i.e. we haven't obviously got much help **in program design** from purely philosophical **considerations**.

YORICK

I'm **not** sure I agree. Arthur, that your two cases are of the same sort, because **Chomsky** et al. have **never been** taken up by AI people in the way the Predicate Calculus was.

Now to Bruce's question of what effect the **construction** of the **all singing, all dancing**, all talking robot would have **on** philosophers. As you suspected earlier it would be zero, **and they'd be quite right**. This ties **up** with my distrust of what Bruce calls "the Philosophy of AI". I've **not seen** that there is **any**, or that there are **any** intellectual **questions** to which AI has contributed a **single** thought (I'd love to be proved wrong **on** this).

A clear proof of this is the endless AI **discussion** starting with the assumption "suppose a robot walked **in** here **and** behaved exactly like... ". Most people who do this don't realise that the **discussion** was conducted **much more elegantly** by Descartes **in** the 16th. Century, **and** that the *nature of the assumption* 'hasn't been altered one scrap by the invention of the coin pu ter.

RICHARD

I **don't agree**. The exact **nature** of the robot would matter. Suppose it had biological **components**, as opposed **to** just digital **ones**. **Or** suppose we could show that a digital robot could **not** exist, i.e. is essentially **unrealizable**. **Certainly** that would alter your ideas about the real world.

## YORICK

Well, **alright**, perhaps I should have said ‘host of the **questions** about the **assumption** haven’t etc....” Bruce, you talk **often** of “philosophical results” **and** their possible **relation** to AI. **and** I think this phrase is near the heart of our mutual **misunderstandings**. For, in a straightforward sense (the one I assume you **mean**) there **are** no philosophical results. And that fact **doesn’t** devalue philosophy **in** the least. It may just be the case that there are AI problems **on the one hand**, **and** there are philosophical problems **on** the other., **and** the two sets are **simply** disjoint. **In** the same way you wouldn’t expect the solution of problems in psychology to solve problems **in** economics. Life is just like that. It doesn’t mean that Philosophy is of **no value** to AI.

Another **thing** that worries **me** is that when AI does come up with a potentially **contentful** general idea, it is hardly ever stated **in** a clear and comprehensible way. Take two cliches of **MIT-AI** “Meaning is procedures” **and** “Hierarchy **not** hierarchy”. There may well be **something** in both of those theses, but I have never seen either of them stated in such a way as to **make** clear that they don’t **mean** the **things** philosophers interpreted **them** as **meaning** at various times **in** the past (**particularly** the first). For, in those **senses** the statements are pretty straightforwardly false. Most straightforward **interpretations** of the first, for example, would **mean** that we could **then no** longer usefully **distinguish** between words whose **meanings** plausibly **are** procedures (like “unscrew”) and words whose **meanings** clearly **aren’t**, like “mud”. **In** the case of the heterarchy thesis, I suspect it’s merely **incoherent**, **and** is **not** a thesis at **all** but a disguised **injunction** to use **certain kinds** of program **control** structures. But of course it’s hinted at that its **much more really**. I suspect that its adherents **haven’t** actually thought out whether they are exhorting people to construct programs in a **given** way, or whether they are **making** a real **claim** about the **things named** in their slogans.

If the basics of such **principles** were set out, the rest of us could get round to **reasoned objections** to what we **think** they **mean** (because of course, while there’s **no** clear **statement** there **can be no** clear objections. Religious leaders have known this for **millenta!!!**)

## ARTHUR

Yes, that’s a very good point. What about the case of the dictum that you **mentioned**: “**meaning** is procedures”? It’s been my **impression** that this was a strong view, well-articulated by **some** of our colleagues: but when I search their papers for that view **neatly** encapsulated, I **can’t** find it clearly set out in anything like that form. Some of this **may be due** to a modest **realisation** on their part that there is **no** pat answer to the question of what meaning or knowledge is, but somehow that doesn’t emerge at all clearly,

Let’s **imagine** for a **moment** that such a creature as a “pure proceduralist” exists, devoted as he is to the **notion** that **all knowledge** is to be **represented** in the form of procedures (**programs** for doing things). Of course, we might argue **among** ourselves whether such a creature **really** exists. I personally **am** inclined to think that, at least in **terms** of the **implications** of what they’ve said and **written**, that most of the MIT people can, to varying degrees be tarred with this brush. But what would be the effect of such a pure proceduralist approach, such as I still think the MIT school has advocated, **on a general** theory of how robots would behave? I take the proceduralists to be saying that **intelligence** consists of

'knowing how' rather than knowing what': now this view is one of the corner-stones of the behaviourist philosophy that was developed by Gilbert Ryle among others, which argues that talk of mental states is illicit, and that only behaviour and dispositions to behaviour are the legitimate concerns of psychology. This view has it seems to me been effectively demolished, and is now regarded as rather old-fashioned. Do the proceduralists have some answer to this, or is it yet another case of AI rediscovering ancient philosophical controversy?

There is perhaps a further point that should be brought out about the proceduralist thesis. I take it for granted that a proceduralist in AI would want to say that he was doing real epistemology, and perhaps even that he wanted to make his theory of knowledge all-encompassing. Now, if he's honest, he will have to accept that a proceduralist view of human beings might well lead to the impossibility of developing a theory of ethics. What I mean is that, if you press for the impossibility of having any declarative information around (any 'maxims' as Ryle would have it) then you seem to leave yourself no grounds for deciding what to do in a given situation, or on which to choose which of two equally 'efficacious' courses of action to follow. This I find rather worrying. Although I'm not not really saying that we should worry about robot ethics at this point, I do feel that there's a problem here for people, like the MIT defenders of proceduralism, who also seem to feel that AI is the study of general mechanisms of knowledge and intelligence, rather than the construction of robot machines.

YORICK

You're right that there are such clear analogies between Ryle and the proceduralists as, for example, both would want to answer a question like "Does Smith know chess?" not in terms of what Smith knows, but in terms of how he performs. And of course Ryle could have said "the meanings of many mental terms are really procedures", and the proceduralist might find himself agreeing.

But I think this agreement would be almost wholly illusory. Ryle as a behaviourist is interested in external behavior and never in internal representation. The proceduralist is the reverse. The illusion of agreement is because Pseudoryle uses "procedure" to mean "external behavior" and the proceduralist uses the same word to mean "internal process" or "how my program runs". Hence we have a classic misunderstanding.

My hunch is that the real intellectual ancestors of the proceduralist are German idealists like Hegel and Fichte (and Marx to some extent) who really thought that the world was what our consciousness constructed. This is very like the proceduralist/model people who talk of "block" as meaning what is manipulated by their procedural model -- the world has no reality to them over and above what their system does with it. Hegel would have felt fairly at home there, though, as a metaphysics, that is demonstrably inadequate for both robots and for a model of ourselves, because real blocks always turn out to have properties over and above such procedural definitions.

ARTHUR

No, Yorick, I don't see that there is any real difference between Ryle and the proceduralists along the lines that you think. It seems to me that Ryle was in a fairly clear sense

**concerned with internal representations:** his point was just that **any discussion of 'mental concepts' must** be couched **in behaviour terms**, rather than **in terms of occult intellectual episodes**.

Perhaps, to move away **from** philosophical historiography, I **can** suggest a way of **escaping** the **Rylean arguments**, which take the **form** of a claim that **any examination** of maxims before **executing some behaviour involves an infinite regress** (**examining** the maxims is itself a piece of behaviour which needs evaluation at **some meta-level, and so on**). The tool for our escape comes, paradoxically **from** the proceduralist's first love -- pattern-matching. If we have a machine which **can** activate procedures **in** a pattern-matching **kind of way**, **then it seems** we have escaped the infinite regress: the very existence of a **certain pattern** of declarative **information** (a set of predicates or whatever) will **generate** the appropriate behaviour. Does this make sense?

TOM

I don't **recognize** the **proceduralist** position as it's being described here. If we are talking **about work** done at MIT, I doubt that there is a **clearly** defined position. The point of **discussion** is whether they **meant to deny** declarative knowledge. They did not. There was a **reaction** against theories which **denied** procedural **information**. What appeared to me was a **real concern** about predicate calculus **formulations of knowledge**, whose **language** was **inadequate, and** whose **information** was jinxed up in a sea of undifferentiated statements, and that **some form of program control** was necessary. There was **no sense** of **denying** declarative **information**, only that **adding** procedures was a powerful **and** simple way of adding **control** while **localizing** the **context** of information.

BRUCE

Yes Arthur, you are trying to read too much **into** what programmers say: of course Hewitt suggested, and Wierograd used, procedures **in an interesting way**. I don't **think** either of them would want to go further **than** that.

ARTHUR

Oh!, but I'd thought all **along** that Hewitt **and Winograd**, like others of their ilk, were **genuinely interested in epistemological problems**. If you're right (and I don't **think** for a **moment** that you are!), **and** their **only** claim is to be "using procedures **in an** interesting way" **then** their work, **excellent** as it is, seems to lose a lot of its **intellectual force and interest**.

RICHARD

But, **Bruce**, **some** people **in AI** as a matter of fact **now do hold positions amounting to** "procedures are the **meaning** of words". Much **programming** effort flows from this general **position** -- and that is true **even though** they **don't want to defend** or discuss them **philosophically**. They have **influenced** people, **some from** outside the **AI community**, to **accept these positions**.

## YORICK

I agree strongly with Richard about the odd, slippery, way in which the "meanings" are procedures\* position is sometimes held, i.e. it's defended until attacked, and then held unquestioningly again. Nothing improves or clarifies by discussion and debate, even though any strong form of the thesis is clearly untrue.

However, I think its espousal has had a very good effect in the campaign against the MIT Linguists: Chomsky and his various schools of disciples. They have floated so far from any conceivable procedures that a little over-emphasis the other way cannot hurt.

## BRUCE

Don't let's confuse a discussion of the correct methodology/philosophy of AI with a discussion of the relevance of philosophical results for AI. I am sure many people agree that much AI has been done sloppily and there is not enough discussion of basic issues before, jumping onto the console. Of course there are interesting technical topics we can discuss here in the "philosophy" or general attitudes of our subject: Can we sharpen our ideas of design criteria for intelligent programs? How sensible are the various approaches that have been/ are being taken? I think we should move on to more specific areas and to the interesting, and generally unstated, general views that people in AI would have to have to justify the work they are doing.

## YORICK

I agree, but before we move on, may I try a better justification for more philosophical care in AI. Intellectual disciplines progress by the dialectic of assertion and critical counter assert ion. AI is very very short on useful and insightful internal criticism. What there is, by and large, is busy people building system at keyboards and screens in isolation. There is tremendous pressure to be positive at all costs (this may be more an American characteristic than an AI one). The Michie-Clowes interchange is one of the few clashes of view in print that I can think of, and very useful it was. Why is there not more of this? Cod knows we need it! It's not as if people in the field don't harbour very hostile views in private - but these are never articulated or made precise. Here I'm sure philosophy could be very therapeutic, bringing out all those aggressions in a satisfying way.

Pat Suppes once argued that what you usually get in AI, in the absence of rational criticism and discussion, is a series of love affairs: people seize on some piece of work every few years and fall in love with it, then later fall out of love with it. Just as in love, and later disillusion, reason plays no part at all.

None of this is meant to be negative, or to prevent work of any sort going on. It suggests that people should be more generally aware and pessimistic, and then push on anyway. Its disastrous to want to stop even those enterprises one is sure are metaphysically mistaken. A clear case is the dispute between Newton and Leibniz: Leibniz argued that Newton's notion of action at a distance was metaphysically incoherent. And, of course, two centuries later he was generally agreed to be right. Even so, it would have been scientifically disastrous in the short run if he had been able to convince Newton of that and to have stopped his work on gravitation!

## TOM

There seems to be an **implicit argument** that AI **could not mean anything to philosophy**, **not that** that bothers me. But philosophy is said to be **concerned only** with **non-contingent** matters. If **that** particular view of philosophy is **interesting**, then it gives up **vast areas** which traditionally were **philosophy** and which are **now** physics, psychology, **physiology**. **Virtually** anything I care about seems to be **contingent**, and particularly what we are capable (in a hardware sense) of seeing, perceiving and representing. If **none of this** experimental epistemology is relevant to philosophy, then what is non -- contingent? It would seem that **only** formal systems are. In our systems we don't capture the world, only our model of **it**. But in some way, philosophy is not allowed to ask whether the model behaves like the world out there, **only** like **any** possible world out there. How do we choose axiom **systems**? I realize that there is a lot to do with formal systems, **but that field is crowded**, what with mathematics and AI both involved. I would like to talk about a practical epistemology for an **intelligent** being. Richard points out that if we could **show** that the **human brain** can be imitated by a finite state machine, that tells us a lot. He also points out that the **presupposition** of AI (and any science) is that an **enormous** part of the **universe** can be **modelled** by some formal system. But does philosophy **not** allow itself to care about whether the formal systems model this **world**, these humans?

## YORICK

I think, Tow, that you have a too jaundiced view of philosophers: certainly there are **still many** who proudly claim to be **concerned only** with what they call second-order questions. That's not the **same** as What you call **non-contingent** matters, because some of those **philosophers** would say they were **concerned** with linguistic usage which obviously is **contingent**.

But we needn't worry about them, because **many** philosophers are interested in AI, and it's a fair bet that **many of the** great philosophers of the past would have **been** very excited by it, as they were by all the philosophical developments of their **own** days. Certainly **no one** here, I think, is trying to prove the total independence of philosophy and AI.

Perhaps we should move **on**, as Bruce hinted, to more specific questions. For instance, it **seems** clear to me that **many** approaches in AI are too deductive, **and** that for many **reasons** this cannot be either a fruitful model of how brains work, or the basis of a sensible **informationprocessing** system. What I mean by the **distinction** between inferences and **deductions** can be illustrated off the cuff by **analogy** with **doing** geometry by proofs or by **deductions**. One does school **geometry** examples by proofs, **written** or drawn, yet **one could do them by deduction in** some powerful language, like **set** theory, in which each step was deductively valid. But that would be **insane**. It would be like reading a book letter by letter instead of simply reading it.

## BRUCE

You are-right, but geometry is a bad example to start with as it **can** be formalized relatively easily **and** the **relation** between **inference** and **deduction** is clear. This is **not** so in general, as you will **no doubt want** to say. That's why mathematics is such a **bad** problem area for AI: the facts that formal deductions exist makes people concentrate on making **deduction**.

checkers arid deduction-erigifies without **thinking** about what' things go **on in** a person's head **when** looking for a proof. **Or even when understanding one** -- it is clear that published proofs (which aren't usually very formal, actually) are only the surface manifestation of something much deeper.

Whatever the procedures **running** in the head are, it seems better to play with computer **models of them** directly rather than with logical descriptions of them. They can be **described in** logic but I don't **think** they **can** be **modelled** by it.

## YORICK

**One** could support this point with **an** analogy from scientific method, where the question of the **axiomatization** of a scientific theory **only** arises **after** there is a theory. **In** many areas of AI people are trying to go directly to the axiomatization when there is **no** substantial theory to axiomatize. They simply assume that the process of axiomatization also, and at the same time, provides some **content**.

## BRUCE

Yes, 1 think the **distinction** between **axiomatizing** a theory in a logic and **modelling** thinking **as** deductions of a logic has escaped several people in the field.

## RICHARD

I don't **think** that the **distinction** you make **between** deductions and proofs and inferences is as **simple** as you pretend. Your **own comments** seem to me to indicate a lack of "mutual" **understanding** of what words should be attached to what notions. I propose that by a proof of some fact (or sentence if facts can be expressed by language) we mean whatever it is that carries conviction for us, i. e. what convinces us that it is true. **Better** yet whatever evidence it is that allows us to assert it as a fact. By a deduction we should mean what people usually call a "formal" proof. These require a language in which it is decidable what is **and** is not a **sentence** (is English in this class ?), together with a decidable predicate  $\text{Prf}(x, A)$ , **which** singles out as assertable those sentences  $A$  for which there is an  $x$  with  $\text{Prf}(x, A)$ . The **nature** of the allowable  $x$ 's or how you discover them is irrelevant, it is the ability to **decide**  $\text{Prf}(x, A)$  for any particular  $x$  **and**  $A$  that makes the "proof" formal.

This **notion** of "formal" proof is of course very wide. It includes all **computations** (and maybe more **depending on** what it **means** to decide). This was intentional, as I wish to **emphasise** that all representation theory **in** AI is caught doing deductions in this **sense**. I **believe** that **my distinction** is more weighty than just saying something like "of course we **are always doing deductions**. all we have are digital computers". Namely it shifts our **emphasis** from arguing over how "formal" your way of doing AI is, to "What is the nature of the formalism that I **am** proposing?". It is **only** bad **propaganda** and sloppy **thinking** that **allows** yourself to be drawn **into arguments** about the "informality" of an approach to AI.

Using this terminology, Yorick, I uriderstafid you to **mean** that you find the traditional deductions, e.g. **in** the lower predicate calculus, or some of the usual forms of set theory, at least as expressed in terms of axioms **and** "deductions" by suitably applying collections of rules of **inference** and theorems, are unsatisfactory as a theory of reasoning.

In light of the **distinctions** I **ain't** trying to maintain I would like to **introduce another notion**, **argument**, (a **term** once suggested by Bruce) to **mean** those **kinds** of things usually **written down in books** which **tend** to **convince** us of **some** facts. This notion is **distinguished from** proofs in that **arguments** are linguistic in nature. I see **arguments** as representing the linguistic traces of 'Proofs. With this **new notion** at hand, (correct me if I'm **wrong** Yorick), you **seem** to suggest that **inferences** in AI should be made in some formalism whose basic **building blocks** look more like **arguments than traditional "deductions"**.

Contrary to you Bruce, I **think** that geometry is a particularly good example. The distinct **ions** I just **mentioned** are clear there. The **arguments given in geometry** texts are **compelling**, that is they **seem** to carry conviction, and thus qualify as proofs, but as it turns out the arguments given in most **secondary** school texts are formally inadequate, in **that the** continuity axiom is **missing, and thus** cannot be justified by deduction from the usual **Euclidean axioms**. It took as good a **mathematician** as **Hilbert** to correctly formalize geometry.

I see the problem **differently**. One question to be asked about geometry is whether or not the **arguments** as presented in **elementary** texts, as the traces of proofs, **can** be generated in a formal way at all. Another, **more** relevant here, is what exactly is the language of **arguments and** what is the **corresponding notion** of valid **consequence** for them. I feel that if **you cannot** say something clear about that **then** you are not talking about AI (which at present involves **digital** computers).

BRUCE

I would like to argue that McCarthy's **distinction** between the "heuristic" and "epistemiological" adequacy of a **reasoning** system causes **some** trouble. Suppose we have a **system in two parts**, the facts and **inference** rules (of course John wants us to have **some** fairly straightforward **logic here**), and secondly some **engine** which decides which inference to do (I **think John** thinks we **can** worry about how this works later). The system is **meant** to be **epistemologically** adequate in that all the right inferences **can** be made in the first **part, and** heuristically **adequate in** that the second part actually gets them **done** appropriately. Let's call these the **axioms** and **strategy** parts. Using the **system** to represent **an agent's** knowledge, clearly the **agent** doesn't know everything that follows from the **axioms, only** those things that the **strategy** "allows" it to deduce. But if I want to talk **about what someone else knows**, my **axioms** must cover his **strategy**. For **example**, the fact that he never does proofs of **more than three steps** must be described. This is going to be quite a system, **and of course the ordinary logics** people **have been** using have nothing to say here **at all**.

ARTHUR

Well, there are two points that I **think should** be made: firstly, I **think** there is **some confusion in** the way that McCarthy has used the **terms** 'epistemiological' and 'heuristic'. He **can** of course use those words to **mean** anything he likes, but it's unfortunate that they already have well-established **meanings** for philosophers -- meanings which don't seem to overlap precisely with his. McCarthy's term 'epistemology' seem to have features which **traditionally have been** regarded as **being metaphysical and ontological**, as well as

epistemological. Traditional epistemology is **concerned** mostly with the actual process of acquiring knowledge. Metaphysics is **concerned** with the limits which are placed **on knowledge, and** ontology of **course** is concerned with the **question** of existence,

But **secondly**, even though I'm unhappy with his use of words, I really feel I must **defend** what I take McCarthy's basic **point** to be: that it is worthwhile exploring the limits **on the expression** of knowledge **independently** of actually trying to express **something** in particular. This I think is a valuable **insight** which deserves stressing to AI types, who are **generally** quite **ignorant** of the fact that this is a well-established concern of philosophy. His notion of 'epistemological adequacy' is, to my mind, extremely **important** if we are to **get** anywhere with the problem of representation. It allows one to say "**aha**! yes I see that I really **need** to model his strategy **in my language**" without having one's head cluttered by **worries** about problem-solving methods, *per se*.

BRUCE

No, **you** have abstracted his **position** to the **level** of remarks such as "**think** carefully", whereas the **argument** is a much **more** technical **one than** that.

RICHARD

Bruce, **what** kind of "reasoning" do you propose that is **not** related to **some** calculus for making deductions? The study of (or **notion** of) the validity of this **reasoning** is surely in the traditional realm of logic.

BRUCE

**Traditional** but **not** modern. Surely **most logicians** today **don't think** they are studying how people think? And if I tried to pass myself off as a logician people would **think** I was joking. **When** I say "logic is **no** good" or something like that, I **mean** that logicians don't have anything to tell me about how people **think, and** their formalisms reflect this. Now if **you want** to say that **any** search for a calculus for **modelling** actual inferences is by definition logic, then tell me why more logicians don't do logic!

RICHARD

I **think** you **underestimate contemporary** logic. Metamathematical studies **and** proof theoretic studies are centrally concerned with the questions of **both what kinds of objects mathematics is about, and** what kinds of **evidence** is acceptable **in** making **inferences**, either using proofs or **in** deductive system. **You** are not clear **when** you say "actual" inferences.

YORICK

Well, OK, Richard, you **want** to use "**argument**" as the word to oppose to "**deduction**", **rather than "inference"** as I suggested initially, and that's **fine** by me -- though I think there's perfectly-good traditional justification for the one I started with.

Your **notion** of "**proof**" is very interesting in itself, but doesn't give us **anything** to really **get our teeth into** as yet (without more work on your part) because by definition it's an

entity existing in a non-symbolic (and not merely non-formal) realm. Perhaps you should tell us a little more about what realm it does exist in? Also, not all our differences here can be cleared up simply by agreeing which words to use and which words to oppose to each other because, for example, you have, I think, a much more formalist idea of deduction than I have -- so for you virtually my formal manipulation is a deduction, whereas for me it has to have some connexion with the sort of thing traditionally meant by deduction, that is following by means of a rule expressing a logical truth (in some irreducible sense of that phrase, What I mean here is something along the lines of what Davidson has expressed recently and very well with his "In defense of convention T"). Is there a real difference here or am I just not seeing something modern and obviously true?

ARTHUR

Sorry, but I want to be boring and go back for a second to what Bruce said earlier about logic, since I feel he really is suffering under some misapprehensions about it. A major part of modern logic is model theory, or formal semantics. And model theory's major concern is with the question of what can and cannot be expressed in a given type of language. Surely that must be a central concern of anyone who is interested in expressing knowledge in any formalism whatever.

Also, I don't think, Yorick, that you're being particularly fair when you say that Richard's notion of 'proof' is unsystematic. People are just now beginning to have some rigorous insights into how people carry out proofs, and I think it will turn out that one can talk about them in a much more substantive way than you think possible.

BRUCE

Arthur, I know you think that, but who are these people? What are these insights? Or is it all just a feeling? Don't get me wrong, I don't object to feelings, but I don't think you should be allowed to get away with saying that "... people are just now beginning to have some rigorous insights into how people carry out proofs..." without some justification. I'm not aware of any results in psychology -- surely you aren't talking about results of logicians?

ARTHUR

As a matter of fact, I am. You know, of course that there has recently been considerable effort by some logicians, of whom Kreisel is the most prominent, to get some systematic insights (from a logical point of view, of course) into the nature of the curious mathematical objects that we call proofs. And, unless I'm much mistaken, the study of the metamathematics of proofs is one of the things that Richard has in mind in his work on a first-order machine proof checker.

BRUCE

I spy an attempted proof by repetition! The "systematic insights" you speak of are of interest to those working in the foundations of mathematics, and to some philosophers: but I want to know about what goes on in people's heads as they become convinced of something. People in foundational studies don't address themselves to that problem, at least not in any direct way, and it isn't even clear that they should!

## YORICK

Arthur, I **wasn't** in any way accusing Richard of being unsystematic. He's putting a **novel** idea, **and claiming** that proofs exist in some non-symbolic realm. I suspect there's a lot in his idea, but even he isn't claiming that its systematisation arises at the **moment**. I was pressing for "metaphysical exposition" of the idea, as it were, and that comes way before **any** formalization of it.

Let **me** propose a naive example of actual, or "**contentful**", inference in natural language analysis of what I call "preference **semantics**" (PS). Suppose we are **analysing** "He pushed the book off the table and it fell". We want to know whether the "it" refers to the book or the table, and we can all see it is really the book. What I think of as the PLANNER or deductive method here would want to use, in some way, a "theorem" of the form "Unsupported objects fall". It would have to find that it was a **relevant** theorem and then put it into some deductive structure together with the representation of the example **sentence**, and perhaps other knowledge. What I call preference semantics would look into what it knew about the meaning of "fall" and see that in its representation it preferred unsupported objects as **fallers**, and then infer from the example that the book was unsupported. **I'm not** pressing the details of this example but opposing two general approaches, one of bringing in facts from a pile, the other from **scrutinising** the meaning representation you have more deeply and using preference rules.

## BRUCE

But I don't see where the "opposition" is here. There is no way to understand the sentence except by reference to knowledge about falling, books **and** tables. Whether the relevant facts are **in** a pile of theorems (which is obviously structured in some way to allow sensible access) or a pile of "meanings" (ditto) is **irrelevant** at the level of our discussion. And we would ask the same question about both implementations. For example, suppose the previous **sentence** were "The book tied to his waist lay **on** the table which was tottering on the brink of the abyss, and was the only thing keeping it in balance.", then how would the system's state have been differlit so that the **pronoun** refererice was **done** correctly?

## YORICK

No, Bruce, of course I am not denying that knowledge is needed to settle such matters: how could **I** be, for what else would settle them? **And** all the **elements in** the example I outlined are **clearly knowledge**. It is true that I **am emphasising** again a distinction I made earlier between fact.5 and meanings. The fact that drinking is **essentially** of liquids is not just a fact -- if you think it is, ask yourself how matters could be otherwise while drink retained it.5 present meaning? Whereas, that hands have 4 fingers *is* a fact, because they might have eight without **changing** the meaning of "finger" or of "hand".

This distinction is important here because to see that questions are about meaning encourages one to see them as structured: the whole "facts" approach is inherently atomic, and leads to the view of piles of unstructured "theorems" which you too are against, I know. That's the opposition. I **know** you **want** to say that facts **can** be structured too -- OK, **and** recent things like **Minsky's** "frames" are indeed attempts to structure facts in the same sort of way (as active slot-filling patterns) as preference semantics tries to for more conceptual objects. But its going to be a hard row to hoe, because of the sheer multiplicity of them.

Now to your example, ok so it would fool **my system** in its basic **form**, because it was designed to do so. And to get it to do that you had to produce a sentence that is simply not **how** that message **would** be conveyed by a **competent** human speaker. You've had to (in **my** terms) satisfy a preference and overthrow it in an awkward way -- and the awkwardness isn't accidental. If you think it is, provide an example that isn't awkward.

For **any** system you can design examples to throw it. **So** what? What is a good system for you?

BRUCE

No, I didn't **think** of your **system** (or **even** need to know its details) to think up **this** example. Its a question of two different pieces of text setting up different expectations. A general idea -- perhaps its use qualifies **me** as a philosopher!

YORICK

**Fine**, but you're talking **now** iii terms of Charniak's system of setting up different expectations **in** advance, with what he calls "**demons**". Mine works backwards and forwards from problem-causing **pronouns**. There's something to be said for both approaches: from **my point** of view I prefer a system that sets up **all** this machinery **only when** it has a **problem** it **can't** solve by simpler methods of inference. The massive forward **inferences** to **no** purpose that the **demons** do seems to me **computationally** hopeless.

But **my point** here is that, for every example of yours that satisfies a preference and then overthrows it, I can set up an example that satisfies a **demon** and then satisfies another one, **inconsistent** with the first. So what, still, is a good system for you? Given your **premisses** you **should** like the example I gave, it seem to me.

BRUCE

Well, I don't realty want to say that Charniak's system is the right one either, but I **certainly** agree he **might** get my funny **example** right. For **him** the problem is resolving **some** conflict **between** the "if **something** is **falling** it could well be the table", fired up by the first **sentence** and "if **something** is **falling** it **could** well be the book" fired by the second. I would argue that here **we** have at least some way of **talking about** and perhaps **in** the program **resolving** the difficulty, whereas PS as you have presented it is too rigid: you **seem** to regard preference as the answer rather **than** just a good heuristic.

YORICK

But Bruce, the method you've proposed doesn't lead to any way of solving the difficulty at all, and as we all **know**, there can be no **general** way of **locating** contradictions. What you're expressing is **an aspiration** that such a **contradiction** will be found. I'm prepared to bet that **in** any system where every **sentence** fires up large numbers of expectations, whether or not a problem demands their firing, **and** so on right through a story, will **never** locate **any** such contradiction at all.

**In** any case there's **no** problem at all **in** my system in **accomodating** a specific overthrow of

a preference, **in** such a way that the **system** knows **something** odd is **going on**, as **in** a case where we are told that a **bottle** is made of steel specifically, **and then** an ambiguous **pronoun reference** problem arises whose solution rests **on not then** applying the preference of "break" for **fragile** **breaking** things, because we **now** know something special **and** odd **about the** bottle, as **in** the **sentence** "He dropped the bottle **on** the table **and it broke**". **There's no problem** there for a system that sees a preference is **being** contradicted **and** keeps that fact around for a while.

**What's most disturbing to me** about your example **and** your discussion of it is that you **don't seem** to see the need for a **system** of local **inference** **iii** natural **language understanding**, as a **pragmatic** fact about the language. That is to say, a system of **local preference** that **can indeed in exceptional** cases be overthrown and be superceded by a system of "global" hacks. I produced an example of such inference, **and** you **seem** to think **that** you're showing something by producing a **clumsy and** complicated **counter-example**. You're **not**. **In fact you're** rather **helping me** make my point, **namely** that **any** theory like yours (following Charuiak) that **thinks** you **can understand** language texts with **only** global **expectations** is computationally hopeless **and** psychologically implausible.

**My precise answer** to your point remains that for every example of yours that requires hacks to supercede preferences, there will be an example of "contradictory demons" requiring similar hacks. But the preference system at least provides a psychologically plausible **theory** of local inference, and the other one doesn't.

BR UCE

Obviously "Charniak" means different things to **different** people. It **seems to me** that a **program which is reading and understanding** text **should build** up a model of what the text **is talking** about, as it is reading, **and** use this to help the **understanding** e.g. to **help find referents of pronouns**. The model would for example keep track of who is where at what **time** (**in** the imaginary world of the story): then uses of the word "he" might have their references decided by using this **information**. (You need syntax too!). A problem **immediately** arises: how do you find the **relevant** parts of the model at **any given time**? **Charniak's idea** was (something like) "let every thing **in the model** look out for text **later in the story** which **might** refer to it". Of course there are problems with having too **many demons and having conflicting demons**: the whole system needs much more structure, and **indeed** Charuiak **didn't** say how to get over these difficulties. **So** I see an attempted solution (**demons**) to a problem (relevance) raised by a theory (**maintaining** a world model). Perhaps **my** dissatisfaction with PS is caused by my inability to make this decomposition for it.

We do need to make "local" inferences, but the measure of locality surely **refers to distance** **in** some complex structure **representing** what went **on** in the (imaginary) world of **the story** and the **importance** of different facts and events. **Indeed** Charuiak had **no such structure**, but you almost deny the need for it, by **substituting** "local in the text" for "nearby in the **model**". Of course this **approximation often** works, otherwise I **am sure you wouldn't use it**, **but** you have **not illuminated** whatever it is an **approximation** of! It is often **necessary** to **distinguish the story from the way the story is told**, for example to deal with flashbacks. You don't do **this** (**fine, nor** does **anyone** else), but you don't see the need either!

YORICK

Well, I can't **make much** of that because I don't see any **content** to your "model" or, "theory" or **even** a running system to back the **notions** indirectly. All I see is an **aspiration** to build something that will somehow "**know everything** about everything". But that's all square one stuff as far as I'm concerned. I was trying to offer a concrete example from a concrete theory **embedded** in a **running** system. What puzzled **me** was why you bothered to attack it so. Why do you always go **on** about texts with puzzles **in** them such as flashbacks or clever overriding of preferences?

BRUCE

I thought you would ask that. I **think** this is where we differ: I am **saying** "think about **these** funny things, they **seem** to exemplify (perhaps **in** **some** extreme way) what goes **on a** lot **in** **natural language**", and you say "actually they hardly ever occur arid I'll worry about **them** later". I think your theory has a hole **in** it, whereas you just think it needs **extending**. Presumably **only** future attempts **on** larger **domains** of discourse will resolve the **argument**.

YORICK

No, I'm pretty sure that's not where we differ because I also like to **emphasise** difficult **things** against the **proponents** of simplistic theories of language. I **think** that I **can** deal with the **things** you **mention** by **extensions** of the mechanisms I propose. Where I think we differ is that I **think** you have **no** theory of **language** (as distinct **from reasoning**) at all, nor do you see **the** need for it. Your **distinction between** the "non-linguistic story" **and** "the way **its told**" makes this clear. What people have to **understand is the** way it's told. And, if it's told **in** certain ways they **won't** understand it' whatever a theory of reasoning may say to **the contrary**.

You, like Minsky and Charniak arid probably **many** more, think you **can** assume some abstract linguistic representation' **not** bother to actually apply it to **language** material, **and then** get on **with the** "interesting" stuff like the "reasoning" **and so on**. This view is **profoundly** mistaken because the possible inferences also *determine the form of the representation itself*. In the simplest cases, possible **inferences determine** which sense of a word is the correct one, **and** hence the form of the **representation** of the sentence containing it.

BRUCE

No, I don't have a theory of **language** in the **sense** of how to string words together. I **happen to think** (with others!) that the "**inference**" bit is what we currently **need** to work **on**, and I was looking at your system as an "**inference**" system, without thinking about how you actually gobbled up text **in** the input. Perhaps you think I **shouldn't** (or really can't) do that.

At least -we agree **on** the deduction/inference dispute. **Any** sort of interesting notion, such as "like" covers so much -- you really can't represent it by a predicate -- that in **any** realistic system a complicated structure of **notions** and inference rules will be needed for it. For example, suppose Fred's **saying** "I like fish" becomes  $\forall x. \text{fish}(x) \Rightarrow \text{likes}(\text{Fred}, x)$ . But we

know the following: he probably doesn't like fish that has **gone** off; he may well **not** like **certain** fish cooked in certain ways; there are probably fish he dislikes but has never tasted. **He may** have **forgotten** he dislikes rock salmon, but we don't interrupt him with these objections, unless we are "logically"-minded pedants. **And** if we did, he would say "Come **on**, **you know** what I **mean**" and **indeed** roe **would**. In other words, we **cannot model someone's** **liking** for fish with the simple **sentence** given above. Now the **more** sophisticated logic types will say they **never intended** such a simplistic representation, but they never say this **unless** pressed, and never **seem** to attempt the fuller **axiomatization**!

YORICK

I agree with you entirely about the **importance** of setting up systems of inference for natural language prior to any attempted **axiomatisation** of them (**something** that's taken for granted in all other sciences). Let **me** just add here that what I said earlier about the "**non-availability** of contradiction" iii general was **meant** to apply to the analysis of stories and texts. I didn't **mean** to deny its value (1) in robots and (2) in simulated model worlds.

In the case of a robot, really **moving** about in the world with deductively manipulated **information** and plans, the world itself provides a **clear sense** of **contradiction** : if all the robots deductions tell it the door is open, but it **bangs** into the firmly closed door in fact, **then** the conclusion is contradicted and the preceding premisses can be **reexamined**, as would be the case with a scientific theory refuted by unsuccessful experiment. That is to say, the **premisses** may be unreliable, but because there is **firm contradiction of conclusions** the deductive machinery can **transfer** "not" back to the premisses by modus **tollendo tollens**.

This **situation** I maintain is quite **different** from the analysis of **continuous natural language** where there is little or **no** expectation of contradiction: if, in **understanding** the text, the **understander** erroneously infers A, there is **little** or **no chance** of encountering the assertion **-A** in the text in the **near** future.

In the case of model worlds, simulated after the fashion of Winograd, something else occurs. Here there is **no** contradiction at all, but there is **no** cause for it since all **premisses** are, in effect, **analytic** and **no** real **information** can ever **enter** the system. For example, after **executing** the **command** "Clear off the top of the red block", it is clear by **definition**. No **lingering** and sticky cigarette end can remain to imperil the stability of the **house of** bricks about to be built. It will be clear that such situations have little to do with the **unreliable** inductive **information** required for the analysis of natural language.

BRUCE

To deal with the latter point first, in the perfect toy world **things** indeed **never** go wrong, but that **can** be a valid **simplification** if some other point is what is at issue. In the robot case, the **contradiction of conclusions** is **not** as firm as **you** think. Take the example of **putting in a bolt**. If it fails, i.e. the bolt is **seen not** to be in the hole at the end of the **attempt**, there are **any number** of possibilities for what went wrong and where the bolt is **now**. You **cannot** possibly afford the strategy of checking each micro-step as you go along either, though of course because the task is governed by the laws of physics there is a good **chance** of eventually **finding** out, whereas with people this **kind** of experimentation is usually impossible. "She said she'd **meet me** outside Lyons at three, but I never saw her again."

YORICK

You're right, about the robot **and** contradiction, of course. I was **only** trying to **make** the point that **in** the **analysis** of texts the role of **contradiction cannot** be central. **One** could **not** just throw **iii** any old rules, as one might for dialog, saying "oh well, if they go wrong then the other participant will let us **know** somehow, that we've **gone** off the rails.

BR IJCE

Oh I **don't** know, what about Agatha Christie **novels**?

YORICK

I don't **understand** why you say that at all!

BRIJCE

The **point** is that **in** mystery stories you **do make** assumptions, sometimes **unconsciously**, and **you are** able to deal with things **when** the facts (of the story) coritadict **your model**. "I was sure the butler did it, but **ther**e was a clever twist at the end."

YORICK

Oh **sure**, **there can** be clever twists at the end, just as there **can** be jokes, **puns**, lies, and poetry. The important **thing** is that **most understanding** is not of such things. This cycles straight back to our earlier point of dispute, where you **think counter** examples knock **down** theories of **normal** inference, whereas they don't, but **only** show the need for supplementary theory or hacks. What you **don't** see is the **need** to put anything **contentful** in the center, **because you seem** to think that every utterance is a puzzle. it isn't. It's **only the** schizophrenic who **wonders** (using all his global knowledge about everything) **whether** the waitress is **propositioning him** when she asks "Can I help you, sir?".

May I add two clarificatory points about what I meant when I referred to PLANNER just **now**. I am **not** *opposing* PLANNER-type approaches to more **conventional** complete **methods** **in** theorem **proving** here. . For me, they are **only interesting** different methodologies, but both aim to set up deductive structures **in** a quite **conventional sense** -- as distinct from PS structures, which for example would tolerate the **coexistence** of, say,  $H(a)$  and  $\forall x. \neg H(x)$  in a way that **no system can and remain** deductive.

BR UCE

You are quite **wrong** about the **kinds of inference** people **want** to do in PLANNER (whatever *that* is!), and I thiuk this is the source of our disagreement. Like you I **don't** find logic or formal **semantics** very useful (or **even** illuminating), so let **me** say a bit about my view of logic.

YORICK

Well, I may well be **wrong** about what they **want** to do, I'm **not** privy to that, but I'm pretty

sure about most of what they've done. It's the old difference **between** what is **and** one's aspirations.

BRUCE

I have argued above that the **inferences** people make in everyday life, **and** which we would **like an intelligent** computer to be able to make, cannot be **modelled in** a straightforward way using a **simple** logic. Firstly the logic would have to be self-referential **in** order to deal with **inferences** about other **people's inferences**. **Secondly, many** attributes **cannot** be described by predicates, **nor** is it clear what the **domain** of their values **would** be if they were described **functionally**. Before mentioning a few **more** difficulties I should say that shooting at "logic" **can be done** at many levels, from a rejection of system based **on** any not **ion** of truth, through dislike of the **current** crop of modal logic ideas **and on down** to sniping at first-order predicate calculus. **Doing** the latter has led **me** to the former ! But the arguments **against** the simpleminded approaches are so overwhelming that it really does surprise me to **find** people still **peddling** first-order **logics**. **Unfortunately** this is **not** a **straw-man**.

Consider what happens **when** you make some decision based **on** what you **know**, but **when** you find out **more** facts you reverse the decision. We **cannot represent** this by:

$A \Rightarrow C$

**and**

$A \wedge B \Rightarrow \neg C$

**since** these are contradictory ! Now it could be (probably is) that what happened was that  $\neg B$  was a hidden **antecedent** of the first **inference**, **hidden in** the sense of being **ignored**. But **clearly** we **cannot in** ally reasonable **system represent all** the **antecedents**, such as "if there **isn't** an earthquake" "if I **don't** have a heart attack", "if relativity **continues** to hold (at **least** approximately!)" and **so on**. This is McCarthy's **qualification** problem.

**Another** difficulty **is that logical implication** does **not** correspond very **well** to the **notions** of causality which it is **often used to represent**. Far too much follows from **finding an inconsistency**!

ARTHUR

**While** it's true that **in** a simple-minded logic the kind of problem that Bruce just described **would** be fatal, I **don't** think there's **any** difficulty **in handling** it **now** that we have **much** better **insights** into the notion of **entailment** than that captured by strict **implication**. Everybody knows that strict **implication** leads to paradoxes of the form

$P \text{ and } \neg P \Rightarrow \text{some } Q \text{ and some } Q \Rightarrow P \text{ or } \neg P$

**It's** also well known that one of the reasons for this is **the interdependence** of truth **and** falsity **in** the classical system. **Systems** of entailment like those of **Ackermann, Anderson and Belnap and so on** **seem** to be able to handle the paradoxes, arid so they remove the problem that worries you that a contradiction implies anything.

How would **you** deal with the **hidden antecedent** problem in PLANNER - surely **you'd have** to have a **demon** on the lookout for the **occurrence** of 'B' and when it was activated what this **demon** would do is **change the procedure** call which had previously **handled**  $A \Rightarrow C$ . This is also the **kind** of thing **one** does in logic. In some **sense** the two **implication** symbols **would** have *different interpretations* - different Jnodes. **Logicians** are actively **working on** this topic, so **one can** hardly **claim** that it has **been** ignored.

BRUCE

I **must** be more careful. I suppose I was **trying** to fire in two **directions** at **once, namely** (1) **sociology of logic** -- why are the sophisticated approaches you advocate **not** the ones **actually being followed up**? Do people **think** that because *some* logic **might** be useful, *all* **logics** are thereby **made interesting** to work **on**? (2) I actually don't **think** the **notion** of **truth** is at all basic.

Of **course** it would be foolish to suggest that **logicians and philosophers** haven't recognized **and worked on** at **least some** of these **problems**. But as I've said, I feel that their results, **in terms of formalisms**, are **not** much use to us. That isn't to say there are **no** useful ideas **in logic**: **on the contrary** the ideas of **quantification**, variables, scope and **binding** used **even in** first order predicate calculus have all **been** incorporated in programming languages, as has the **notion** of possible world **from** modal logic. And of course the way logic allows **an axiomatization** to be built up **incrementally**, with the various **sentences being independent** is **something** that **designers** of languages for AI **systems** **designers** strive to allow. However **in languages** such as (the mythical) PLANNER there are powerful **computational** devices available which allow **many more kinds** of inference: interrupts, parallel processes, demons, monitors, sharing, programs as data.

ARTHUR

Ah, but there's the problem that I've tried to point out to you in previous conversations, Bruce: the **problem** that **neither** PLANNER nor its **descendants**, all of which have the **notion** of possible world, **honestly faces up** to the ontological issues which arise. This is the **problem of individuation** -- there seems to be **no** facility **in** these languages to **handle** the question of how to **make identifications between** individuals **in one** possible world **and the same or counterpart** individuals **in** other worlds. PLANNER may appear **on** the surface to **handle** the traditional **problems** of failure of substitutivity of **equivalents** and **existential generalization**, but **on closer analysis** we **find** that it's **in fact** evaded the really hard issues **completely**, by having **dummy** variables which **cannot** be **identified** across **contexts**. So **one** has to be pretty wary of **saying** 'Oh, PLANNER and so on have coped with all the **logical/ontological problems**, and they give all the extra goodies to boot'. They may do the latter, but I **remain firmly** skeptical that they have **done** the former.

BRUCE

Here is a **very simple** approach to the "Bill likes fish" **statement mentioned** above, **Don't** take it too literally -- at this level **QA4** and **POPCORN** are **indistinguishable**!

Bill likes fish.

TO-INFER [Bill likes ?x]

**then** INFER [fish ?x]

This suffers **from** all the problems of the PC representation, but for example **adding**

Nobody likes **mouldy** things.

WHEN-INFERRING [?x likes ?y]  
**then** (INFER [tnouldy ?y]  
→**FAIL inference**)

**now** stops the “Bill therefore likes **mouldy** fish” mistake. Now of course in both systems the **original** rule could have **been changed**, but the **point** here is that **in** the PLANNER **system** we **could** add the rule about **mould** *later and separately* **and** get the right answer. Suppose we **decide** that Welsh people like **mouldy** fish (but **not** anything else mouldy), then

The Welsh like **mouldy** fish

WHEN-INFERRING [?x likes ?y]  
**then** (INFER-SET ([Welsh ?x], [fish ?y][**mouldy** ?y])  
→**/ignore all more general inference monitors/**)

will do the trick. Of course there will be **objections** to this, but they will be **mostly** to **details**, to the actual **representation**, i.e. **arguments about liking** rather than about **schemas or logics**. Well perhaps there is **one** general objection -- “You **aren't using** a well-defined logic so how do you **know** your **system** isn't **inconsistent**?” A quick reply is that this is a **universal** problem for large systems, or **even** for small **ones** **judging** by the **number** of **inconsistent axiomatizations** of Michie's trivial “**Blind** hand problem” that I've **seen**! But a **better answer** has two parts: firstly we **won't** lose as badly as first order logic **because our notion** of implication is much **more causal and** constructive, **and secondly** we have powerful **debugging tools** (tracing, advising etc) to explore and remedy the problem, **so that in the course of experiments we can trace inconsistencies**, perhaps finding some **general class and implementing a solution** with a new piece of **information**.

**Proponents of the logic approach** **may** say that they **can** do all these **things** too, with advice attached to axioms etc, but as we have pointed out above there is a **strong distinction between** advice which speeds up **certain inferences and** advice which **prevents certain deductions from** being **made** i.e. which alters the **semantics** of the system.

RICHARD

I **believe that on both points you are wrong**. This type of rule might be more causal but **certainly** not more constructive in the usual logician's **sense** of the word. To begin with, **constructive** rules are supposed to present themselves as valid. Secondly **these “tools”** for remedying **inconsistencies** **simply do not exist**! Your casual **reference** to “**implementing a solution with a new piece of information**” simply points out that the formalisms you suggest **might sound** good, but **in actual fact** reveals that these formalisms, like the

traditional ones, also suffer from the lack of sufficient reflexiveness or at least our ability to use them in that way to generate programs of their own.

## YORICK

This stuff of Bruce's all seems a good thing to want to do and I'd just like to point out that we have just such unquantified inferences actually running in our system. If we turned your example into a linguistic problem (that's a matter of taste and interest I suppose -- but I feel happier when a thing is not just answering little questions like those of your example) we might have:

"Bill likes fish. The ham is good but the fish is mouldy. Bill likes it."

Our set up would get the "it" as meaning the ham, despite the first sentence, provided we had an inference rule that could be written as follows (with English words for the pieces of semantic coding and numbers for the variables):

(1 BE MOULDY) → ((\*ANI 2) NOTLIKE 1)

where \*ANI simply expresses a matching restriction on variable 2 that anything fitting it must be animate.

## ARTHUR

Perhaps we should move on to another major question: that of 'meaning'. This is of course a topic that is closely connected with inference, and the question of what kinds of actual entities a natural language analysis program should be able to manipulate.

It seems to me that no-one is trying to deny that any significant language understanding system, be it natural or artificial, can get along without a dictionary in some sense. If we are to avoid an infinite regress, the question is rather how we are to define the "primitives" of this dictionary. The meanings of words like 'democracy', as Yorick points out, are not themselves facts, but on the other hand, the dictionary entry for such an abstract word just surely, at some remove, refer back to 'real' facts.

So one might argue that a good way to start developing a formal semantical theory for natural language might be to start with an elementary referential theory and then see how it can be expanded to account for more indirect kinds of referentiality. This, it seems to me, is precisely the kind of thing that has been done recently by logicians like Scott, Montague and Gabbay. This work was aimed at developing a way to deal with the very simplest meaning constructs - those that make direct reference to real-world(physical, geometric) concepts. The work of the developmental linguists (Bierwisch, Clark and others) shows conclusively that perceptual entities are the earliest linguistic primitives that a child acquires. Nobody can deny that as he matures the child uses this primitive referential semantics to construct a more connotative system. Isn't it a bit like the way Ludwig Wittgenstein saw things : the primitive structures, the "pictures of facts" show their meaning directly, while the complex sentences constructed from them only say their meaning indirectly?

YORICK

Well, I can **and** 'do **deny** that **claim** that you preface with "surely" **there**, Arthur. I **want** to reply **along** two **lines**: first, **even** if the **referential constructions** **you** speak of could be **done**, I **don't see** **how** they would **provide** a **form** of **information** **for** a **symbol processing system** **concerned** with **natural** **language**: secondly, the **metaphysics** behind the intended **constructions** seems to me misguided, because words just don't "refer to **things**" **in** the way you **assume** **iii** **e**. As to **Wittgenstein**, **remember** that he **begins** his **best-known** **work** by quoting **Augustine**'s "meanings are thing<sup>5</sup> pointed at" **view** **and then** saying: "**Augustine** describes the learning of **human** **language** as if the **child** came **into** a **strange** **country** **and** **did not** **understand** the **language** of the **country**: that is, as if the **child** already had a **language** only **riot** this **one**." (**Philosophical Investigations**, §32) **In** other words, **referential explanation** is **only OK** if you **know** the meaning of the word **already**.

He goes **on**: "For a large class of cases -- though **not** for all -- in which we **employ** the word "meaning" it **can** be defined thus: the **meaning** of a word is its use **in** the **language**. And the **meaning** of a **name** is *sometimes* (my italics) **explained** by pointing to its **bearer**. "(**Phil. Inv.**, §43) The last **sentence** **isn't even** a **referential semantics** **doctrine** for the smaller class of cases, because he says the **meaning** is *explained* by **pointing** to etc. He **never** says that **IS** the **meaning**. As I **understand** **referential semantics** it says (1) the **bearer** is the **meaning** in general, and (2) hence **Wittgenstein's** smaller **class** (the one in which the **bearer** **seems** at least **relevant** to questions about **meaning**) is really the larger class.

Of course I'm quoting **Wittgenstein** here **only** to **contradict** Arthur **and** to show that he **can** also be quoted against a **referential** **view**.

TOM

If as you say, **Wittgenstein's arguments** **may** be **taken on** both sides, **then** I don't see what **arguments** we really have to suggest that the **referential explanation** is **OK** **only** if we **know** the **meaning** **already**, etc.

YORICK

Oh, **that's** **easy**, we have the **arguments** **he** **put** **when he** was **arguing** **on** that **side**. No **problem** at all. The **reason** **he** **can** be quoted **on both** **sides** is that, like a lot of people, his **viewpoint** **changed and** **developed**. As always, **consistency** **isn't** a **great** **virtue**, **in** **people** **or** **systems**.

ARTHUR

I **don't** **think** that I disagree with the **spirit** of what Yorick says about the **pointed-to-object** **being** **the** **explanation** of the **meaning**, as opposed to **being** the **meaning** **per se**. But **my** **agreement** is **predicated** **upon** **my** **assumption** that there **really** **is** **no** **difference** **between** **explanation** **and** **meaning** -- consider the case of air **electron**. Actually when we refer to "an **electron**" we are referring to its place holder in our atomic theory, rather than to any **coiicrete** **entity**. Indeed there **seems** to be considerable doubt as to whether we **can** ever "know" an **electron** directly. My **claim** is that the **same** is true for all **kinds** of other **individual terms** occurring **iii** the **language**: after all, there are **plausible** **arguments** for

believing that the objects of our direct **acquaintance** are always in a sense **only** ghosts of what we take to be the "real", "concrete" objects. I would **want** to say that the objects that we **know** directly are **models** of their real-world counterparts. I **am** using the term "model" in a fairly strict way here to **mean an individual** within the domain of **interpretation** of the **formal** language describing **my** beliefs: an individual which has properties isomorphic to the "real thing".

This makes **an interesting connection** with the general issue of what **kinds** of things do we **understand models** to be: can we agree **on** some **standard interpretation** of this term?: of what **heuristic** value **can** it be in AI? I would be very interested to get Tom's views **on** this - when he talks about models, what precisely does he mean. This might be a **point** that we **can** defer for fuller discussion till a little later. The role of language (obviously) is to convey **information** about the state of the world to the hearer. The **information** conveyed to the receiver serves to restrict the alternative states of affairs which could exist at that **time** -- it **can** be interpreted by the receiver only with reference to his **own** model of the world around **him**. Jaakko **Hintikka** has made the point that when we are dealing with **quantified sentences**, we cannot **in** a straightforward way compare them to reality **in** the way that we can in the case of atomic **sentences**, as **Wittgenstein** seemed to think. Instead we **must** attempt to construct a **model** in which the sentences can be **imbedded, and** compare these models to reality.

## YORICK

OK, we **can now** drop the metaphysics of **meaning**, I think, because I **now** see that I've **misunderstood** your position all **along**. If you agree with me about **Wittgenstein on Augustine**, and you think "models" are the real objects of reference, **then** you don't hold a **denotational-referential** view at **all**, i. e., that words **mean** real, **hard** objects "out there", **What** I think you should **now** do is explain how what you want is **consistent** with, say, **Montague's expressions of meaning in terms** of set-theoretic **expressions ranging** over real entities **in** the world.

A related **issue** here, about models, is the **distinction between** meanings **and** facts **on** which you touched at the beginning with "democracy". I think **any** sensible system needs this **common-sense** and rough **distinction in** some form, but it is hard to work **into** either a **denotational** or a **model** view. For example, part of the meaning of "water" is that it is liquid, but it's a fact about it that it freezes. Why? Because **many Swahili** speakers, say, know the **meaning** of "maji" but have never **seen** ice. It would be absurd to **conclude** that their ignorance about ice is **ignorance** about the **meaning** of "maji".

## ARTHUR

**While I maintain** that our models of the outside world are epistemologically **and** ontologically prior to what we might call the "real objects", we must **imbed** this **in** a "hypothetice-deductive" framework. **Wittgenstein**, in the "Tractatus" (which of course is the **basic source of my belief** that his **view** was a referentialist one, in that he says something like "The elements of the picture stand, **in** the picture for the **objects**"), seemed to think that we **could** just lay our **language against** the world like a ruler -- I **want** to say that this is **not** possible in **complicated** cases. No, what we have to do is to construct a **theory (whose individual terms are models of things)** and compare that in the traditional common-sense/scientific way **against** the **information** that our sensors give us.

I think that I'm able to maintain a distinction between the individuals in my theory, which I've called models, and what are commonly called concepts. Concepts are much higher-level things, and are more like theories, whereas models are isomorphic to what people think exist in the outside world.

### RICHARD

Yes, Arthur, but it's not as simple as that, because one's notion of validity depends importantly on one's ontology. Tarski's notion of validity is for, and is only for, set-theoretic structures. Now the structures in your model of the world may be like that, and that theory you mention may apply to it. But the real world may not have structure and so the theory may not apply. One can't just say that such logics as Montague's are completely independent of worlds they apply to, or your model of the world. You may actually have to decide whether you refer to models or to real things.

### ARTHUR

Of course you're right, Richard, but I don't believe that I've ever said otherwise: I certainly believe that our models have ontologic priority for us, and that arty interaction we have with the "real" world (assuming that we have it) is strongly mediated via the models. So any logical semantics should, ipso facto, concern itself more with the structure of models than with the structure of the world. And it's not absolutely clear, is it, that the kinds of logical semantics I've been talking about have standard Tarskian model theories: they are much more truth-functional in nature, involving in some sense the so-called 'substitutional' interpretation of the quantificational calculus.

### TOM

Surely, one can't be serious in thinking that there is a meaning for water for us who know ice, without the knowledge that water transforms into ice at low enough temperature? And even without thermometers, we have a sense of what low enough is. Surely, the meaning of water must also change depending on geographical accidents, historical accidents, and the state of one's own ignorance. The fact that ice is not in the Swahili experience would suggest to me that their meaning for "Maji" does not include ice. It would not suggest that the meaning of water is something which is common to all human experience, a least common denominator.

### BRUCE

To be a bit more specific: a person has in his head knowledge both about water and about the word "water" or "maji" or "dwr" or whatever. I think the connection is a fairly straightforward one. Now the Swahili doesn't have a representation of the fact that water can become solid, so some (correct) uses of the word "maji" will confuse him, and others will give him new knowledge about water. Similarly, seeing ice for the first time could be confusing or illuminating. We can see how to make a program act correctly (i.e. like a person) here: are you worrying about the "meaning" of "water" as robot-builders/person-theorists, or as philosophers?

## TOM

I want to know, Yorick, what you **mean** by meaning. My **own** sense of meaning is that **an orange** is **an orange**. The **meaning** of a particular **orange** is the **orange** itself. Now we **can't** possibly keep **an orange** in our head, **so** that we have a **structure** of descriptors (**always** with the **possibility** of referring back to the **original** or a **specimen** of the **class** (go buy **an orange**) to enrich the description) **but** the **important part of meaning** is the **reference** to **examples**. Also, it **must** encompass the possible **experience** with that object or **class**. In **some** cases, that **may** be a considerable body of **knowledge**, and that is the **meaning** of the word, concept, or whatever. What **can** we possibly have but facts (in the broad sense of relations **and references among** concepts)? Thus a **meaning** is a model, which we **can** change by **reference** to the **real** world (**experiment**): of course that **reference** depends on **models**.

I would entirely agree with Arthur's statement that the objects that we **know** directly **are models** of their real-world counterpart. That is their purpose. Language is nothing but a low quality link from one's **models** to another's. We have certain descriptive **elements** and certain **modelling elements** (assume, suppose, and all the imperatives used in that **sense**). People's **models** are **not** at all **identical**, but there **is** something in **common**. Most of my **discussion** is **centered**, however, on what **structure** models should have. Apologies to Yorick on the use of the word **model**. I am equally appalled by his **devaluation** of the word theory to a little word which is applicable to the products of sociologists and ambitious engineers. read on. But what I **mean** by **model** is a structure, since **knowledge** is tightly interwoven. Add to that, **computational structure**, since what I want to do is **compute**. Amend that to read, **computational structure** which **mimics** the world, since what I want to compute is: **given an identification between some models and observables of the real world, can I explain the changes of state of selected observables by their connections with my models**.

Later. I shall want to say quite a bit more about what I take the actual **nature** of models to be; and also to talk about the **relationship** of my ideas to the **nature** of learning.

## YORICK

But took here, you can't get away with this, Tom. You've just **said** that the **meaning** of "orange" is **an orange**. You've also said that we **know only** models directly and they are what words **mean**. These two views are quite **different**, and **incompatible** to boot. You and Arthur **realy** have got to **make up** your **minds** which view you hold. Again, you can't say we **know models and only** models directly, and then talk about comparing **models and** the real world. If your first **assumption** is correct, the **second** task is impossible. Lastly, it's clear we **don't know just** models directly -- from all Tom says about them I seem to **know nothing** about them, directly or indirectly. I feel on much surer **ground in saying** that I **know my own foot directly than in saying** that I **know** a Tom-model directly, and who could blame me?

What do I mean by meaning? Tom asks. When someone asks me the meaning of X I give explanations till he's happy or shuts up. I rarely if ever point to anything; I often refer to dictionaries because what they contain (not pictures in British dictionaries) maps more or less onto what I mean by explanations. From some of the things he wrote I think Tom accepts this **explanation view of meaning**. But, I would argue strongly, the view is *prima*

*facie different from* two other views of meaning he also seems to accept, as I pointed out earlier. I do indeed mean what Tom feared about the meaning of "water"---if we found a substance like water in all respects except that it didn't freeze at all, we would still call it water wouldn't we (doubt that!) - arid probably add "tricky water" or something. It would still be water wouldn't it -- that's my point exactly.

On Tom's general point, about models and understanding, I think there may be no real dispute, only a difference of emphasis, between us here. He wants to emphasise the role of facts in the understander/model more than I do. I want to emphasise conceptual/analytic knowledge. Tom says meaning=model: I say meaning=explanations. There may be no real difference here except that "model" suggests the structure of the explanations is known. I do not think it is known, by Tom or anyone else: so I see "model=explanations" t aspirations ". The aspirational mode may raise the morale of the troops, but I don't see it does anything over and above that. All our positions in this dispute are, I suspect, circular. Tom says that people can translate those things for which they have adequate model systems, in his sense of those words, and cannot those for which they don't etc. I cannot see that that is any more than a partial definition of what Tom means by the phrase "adequate...etc", the whole thing is circular because if a system set up by Badman translated without having some set of facts that Tom formerly considered essential for an "adequate model system", he would then say, oh well, so in this case only a part of the real "adequate model system" was required, but watch out next time Badman ! There's no dispute here about the need for knowledge to understand: only a question of how much and how to organize it, and how to extend it where necessary in the fact of awkward facts.

It's clear that I think that a lot less knowledge will get you along with translation-understanding than Tom does: moreover I think it should be largely, though not wholly, conceptual understanding and not knowledge of superficial facts. Moreover, as I've said before, I think it should be organized nondeductively and have precise suggestions for how to do that. When Tom talks of "models" as structures of facts/explanations, I do not know what structures he has in mind. I really don't. I know to some extent what systems Tom considers inadequate, but not what would do better for him.

## BRUCE

A difficulty/misunderstanding here is that so far facts represented in programs have usually been at a very concrete level i.e. the "Block A is on block B." sort of thing. Clearly one can translate by understanding at a more general level. Putting "The magnet deflected the electron beam" into French could be done without knowing the meaning (or at least the full meaning) of "magnet" or "electron" or "beam": the structure at the level of <actor><action><acted-upon> would be sufficient. In fact, nobody knows the "full" meaning of "electron", at least in some reasonable sense. But please don't let's get into philosophy of science!

## YORICK

No that's not true, there are running programs representing far more complex facts. The point about translation is correct, but says nothing specific about it vis a vis other forms of understanding. People talk quite meaningfully and adequately about magnets in everyday life without knowing much of what you would call the "full meaning of magnetism". I

think you've succumbed to the **AI** mythology that you can't **talk about** anything **properly without knowing all** about it. But, just look at **us**, and most of the world's **population!!**

BRUCE

I am not aware of any programs that really *understand* these more complex facts: I don't mean deep technical **understanding**, but at some level reasonable for a **person**. An ordinary person **talks about** magnets in terms of certain materials, forces and effects. A program that can't do that will not fare well in translating sentences about magnets!

YORICK

Your last **assertion** can be tested quite easily, and **only** modesty **prevents me increasing your awareness in the course of this discussion**.

BRUCE

Your last **assertion** can be tested quite easily, and **only** modesty **prevents me increasing your awareness in the course of this discussion**!

YORICK

To go back a little, I couldn't understand Tom's sentence "The meaning of a particular orange is the orange itself". I don't think oranges have meaning -- except possibly and derivatively as symbols in Prokofieff operas. Word strings have meanings, and I believe those meanings are always other words -- c. f. Quine and Wittgenstein *passim* on "the inscrutability of reference". It is a profound and enduring myth that we mean by pointing --we can never do that in fact, at least not unambiguously, and without the whole weight of the meaning being carried by the language and assumptions we share. See Quine *ad nauseam* on trying to know what a savage is pointing to as he says "Gavagai". Tom sometimes seems to admit this when, for example, he says that "we know only models directly".

The value of the 'meaning is facts view' depends how widely you take 'facts'. Much of meaning is explained by sentences like "Meanings are what words refer to" (false in this case), and "Fascism is the last stage of monopoly capitalism" (false again). But those are not facts, in the ordinary sense of that word. Anyone who thinks they are should then ask himself how he would check up on their truth or falsehood. Most sentences on this file are of this sort. Arthur thinks that all such sentences are ultimately *reducible to* elementary facts. An interesting thesis, but it is a philosophical position, and not self-evidently true. In the common sense sense of "fact" those sentences are not facts. If by "fact" Tom simply means "any assertoric sentence", then ok, but, as he would say, so what?!

ARTHUR

Now I am, and I suspect Tom is, completely confused by what Yorick means (if you'll forgive the nasty word) by "explanation". I get the impression, I hope incorrectly, that for Yorick "explanations" are a never-ending regression of (possibly recursive) pseudo-explanations. You are quite right in saying, Yorick, that my thesis about ultimate

reducibility to **elementary** referring locutions is a **philosophical** position. Surely all of us in **this discussion** are putting forward philosophical positions. We **may** care to support these with **empirical evidence**, but that is in a sense peripheral, in that philosophical **argument** is in essence analytic, or at least a priori synthetic.

**Surely** objects *do* have meaning in and of themselves in a very direct and crucial way. If I say to you "you are sitting on a bomb which is about to explode", the word "bomb" in this sentence is in itself insignificant -- it is the *actual bomb* which is about to blow you to smithereens. As the old saying goes "sticks and stones may break my bones, but names call never hurt me". **Words only** act as *pointers* to the objects which themselves have significance in my life. Tom, and I, would say that their pointing is mediated by Iiodeis for the real thing (without saying that that is all a Inodei is).

**But look, Yorick**, at the very beginning I said that **my** view was that tile lexical entries for **abstract** words referred **back at some remove** to 'primitive' terms which **gain their meanings** by referring directly. So I certainty wouldn't **want** to argue that your "monopoly capitalism" &c., sentences refer directly to some objects. **One** Inight put forward the proposition that if they *did* refer directly to something, it would be to people's behaviour in a capitalist/fascist society, and that this behaviour was captured for us by **intellectual/historical** models which would allow us to predict behaviour under such a regime. I'm not **sure** that I **want** to do that, but it **sounds** plausible if the first approach worl't wash.

Obviously a central tenet of the Iiodei-theoretic approach to language is the **notion of truth**. It is a **weakness in my present position** that I **can't** decide whether a truly **semantical** approach, which **can** be made to work for declarative **sentences**, can be extended to deal with **interrogatives, greetings, commands** etc. That is to say, I'm **not** sure whether the **semantic** theory **can** be made a pragmatic theory. This is precisely what **Montague** tried to do: indeed he went much further, in that he was trying to evolve a **general** theory which **would embrace intensionality, modality and tense**. It's an open question **whether he succeeded but my feeling** is that he set out **along** the right road, and it's up to others to try making the extensions. C. L. **Hamblin** has tried, with some success, I **think**, to **extend** Montague's **notions** to questions.

## BRUCE

**Just** a minute ! There is **this** dispute as to whether tile **notion** of truth is a useful basis for a theory of **meaning**, and though it 'is well known I think it is worth a brief **mention**. Take for example the concept of tallness. We can't really think of it as a predicate (of **one** argument). The question "Is so-and-so tall?" (or even "Do you think so-and-so is tall?") is **not** always expected to have a **one-word** answer "yes" or "no". Fuzzy or multi-valued logics don't really do the trick, as they merely extend the range of **answers** (to "rather", "somewhat" etc) whereas we should really recognize that all "answer" might well involve **asking further questions** as to the questioner's **intention**.

## ARTHUR

**But, that's** precisely the kind of thing that a good pragmatic theory **would** capture for you. Nobody **would** argue that a purely truth-based theory of **meaning** would be adequate in that

*sense.* The claims that have **been** made are two-fold: firstly, that developing a theory of truth is a good (perhaps the **only**) place to start in developing a theory of meaning, **and** secondly, that the central tenet of the theory is more to do with whether the hearer can **imagine** a state of affairs in which a sentence is true. Surely that's sensible, independently of whether the hearer is a **pigmy** or **giant** (each would have a pretty good idea of what the sentence "X is tail" would **mean** to *him*)?

BRUCE

I **don't** want to iabour this **well-known** argument too much, but **unless** you give **me some** more details of what the **pragmatics** has to do, and how you **will** handle the "imagining" you **mention**, I **wiii remain** unconvinced.

ARTHUR

**Yorick, you mentioned** earlier the very important **arguments** that **Quine** and the later **Wittgenstein**, among others, have put forward against a naive **referentialism**; **arguments** which are based essentially **on** the difficulty of **knowing** what it is the **native** speaker is actually "**pointing at**" **when** he utters a **new** word. This is a problem, but actually, of course, it's a **problem** that *really* crops **up** in a child's attempts to **learn** its *first* language. **Piaget** and a great **many** other **developmental** linguists have noticed that kids quite often make the mistake of "**overextension**". That is to say, the use words, say, like "brother", to refer quite generally to all **young males**. It is **only** somewhat later in their **iriguistic** **experience** that they acquire the **distinguishing** lexical markers which prevent this mistake **and allow them** to restrict the **meaning** that they attach to such words. Eve Clark, **Manfred Bierwisch** and others have suggested very **interestingly** that these mistakes have their origin **in the** failure of perceptual discrimination **on** the part of the child. **In fact**, they've gone **much** further along the road of **claiming** that perceptual processes have profouid effects **on** the **development** of semantics by children. **In** the 'Brown Book', **Wittgenstein** seems to want to say that there is no correlation between **explanation** and understanding. He tried to say that what is **involved** in coming to **know** the **meanings** of words is **not** understanding but training. This is surely wrong, in the **same** way as **general** statistical **learning** is wrong compared to learning descriptions. The **referential** view of language **stresses** that what is crucial to **comprehending** the **meanings** of **sntences** is the extraction of the **concepts** behind examples. Early **in** life, children don't **seem** to be able to perform this extraction,

YORICK

By **common-sense** explanations I **meant** the **ordinary** language sense of that word -- i.e. more or less what is **found in dictionaries**. And if you think that's a joke, ask yourself how **you explain** to someone what a word **means** except by paraphrasing what's in a dictionary. Our **task in AI** ought to be to try 'and express such stuff **formally**'.

I still feel that to say that **an object has meaning** is to make a joke or **pun**, and I still find the **last sentence** about **models inscrutable**, and as I've said before, **inconsistent** with the "words **refer to objects(physical)**" view. I still think a lot of quite low level clearing up has to go **into** the Tom-Arthur view before its comprehensible,

To give a **new** twist: to be precise, if words refer to physical object, they do **not** refer to

models, whatever "models" may be. If words are **models** (last sentence I think), then words do not refer to **them(selves?)**, moreover I don't think you can really have meant that, **because**, **Arthur**, I already know you think **models** are set theoretic constructs.

I am not saying there are **not** proper **arguments** about the truth of such **statements**: they go **on all** the time. What is clear is that such **justification** procedures **are not behavioral and** riot proof or set theoretic **in any sense** at all. You've got quite a bit of **justifying** of your **position** to do, Arthur. Let's see an example of a set-theoretic structure: let's see whether it **really expresses** the **meanings of** the words it **refers to**. Your position, like Tom's, **may** just be saying "there is a structured **understanding** system we could build." No-one is going to disagree **with that**: but let's see some **definite content** and above all some **defence** of reducibility to set-theoretic **entities**. We all want to construct **understanding systems** (models if you will). What I've been **objecting** to throughout is the dressing **up of this** enterprise **in metaphysical** clothes which are **indefensible, unnecessary, mutually incompatible and** out of style: for example, you hold that we refer to models when we speak, that these **models** are set-theoretic structures and that they, in fact, are the **only things that** we have know directly. You are, of course, also **advancing** the claim that higher level statements, e. g., about **Germany**, are reducible **in meaning** to lower level **ones about Germans**. None of these **claims** is obvious or necessary to our work. **Discussion of them** is better left to philosophers. It's **no** good **any** of you pretending to despise **philosophy and then full-bloodedly defending** one of these assertions as if it was the merest **common sense**, Tom.

TOM

Of course, we cannot point to a thing. We can suggest a set of **experiments** (look, touch, listen) which have **reproducible** results and we **can** store those results in a **coherent** way. To a **certain** extent we call **communicate** the **results of experiments** to other people, but **not** very accurately. At best we say "Look, now, there. Forever after, I **mean something** like **that when f say orange**." There are **enormous** difficulties **in saying** "similar", but we share the **same meaning of similar**, so **saying** "forever after" works. Let me repeat "a model is a **computational structure which mimics the world**", a model is intended to allow thought **experiments**. Thought **experiments** are safer, quicker, and more **economical than blundering along**. I do not understand what Yorick **means** when he says **explanation**. Does it **mean** substituting word **strings** for word **strings**? If so, how does the poor **soul** who receives this **treatment** **know** when to shut **up**, i.e. be satisfied with the **explanation** (if he **perceives** that is happening, he **should** shut **up** immediately, of course, and seek better company). And do you **deny** entirely reference in language? Do I or do I not have a **structure** "that **orange**" which refers to the particular orange which I have brought **for lunch**? There are two **argumentative copouts** to avoid:

**not** all things are referential:

**reference cannot be infallible.**

Clearly, many words or phrases do not refer to **models of objects**; they refer to models, which refer to other models. Some models have **reference** to objects. A robot must have a **belief in** objects in the world (in the above **sense**). This might seem just a useful **self-delusion**. It seems more **fundamental** in that it **corresponds** to a **discrimination about**

validity of types of knowledge: touch/pain **and manipulation seem** primal-they do not **seem** equivalent to **vision**, hearing, etc. **Solipsism** does **not seem** very important; for a **solipsist** to **function** in a world like **mine**, he **must** use **descriptions** rather like **mine**, with all that **entails**. So reference is **not** infallible, but it is **consistent and predictable** (**I** actually **can** **find** that **orange**).

## YORICK

**Tom**, you say that clearly **many** words **and** phrases refer not to objects but to models -- sorry, but that isn't at all clear to the majority of the **human** race, **including** Jnost of its best informed members. It **sounds** just like **non-common-sense** jargon. Why do things have to **refer** at all to be **meaningful** -- why, why, why? They are all right as they are, you see.

As to **explanations**, yes I really do **mean, explanations**. Moreover, you rightly say, **on** that view, flow could the **one** explained to ever have a definite shut-up-point -- **and** you're dead right, he **doesn't, and** the belief that there is such a point (a definite, logical, **satisfaction point**, as it were) is **utterly** wrong. **Explanations** can **always** be pushed **on** further -- look at this file -- there is never a determinate stop **point** -- just as a **painter** always **could** put another stroke **on** a painting, but at **some** point he merely stops, so with **explanations**. There is always the possibility of **an infinite** regress, which does **not** stop with **any** first **terms** or principles, because **ictionaries** are ultimately circular.

**Can one** press you a bit to say **more** clearly what you **mean** by a "Jnode of the world" -- it's clearly fundamental to your view of AI yet also **doesn't correspond** particularly with **any** of the **standard** **senses** of "model". **In** particular, could you tell **us** how your use of the word 'model' **compares** with the logicians? I think discussing this could be very important.

## TOM

**Alright**, I suppose I should say a bit more about what I **mean** by models in the AI sense, **and** how that **compares** with the **notion** of Jnode used by logicians. **In** logic, a theory is the set of axioms (e.g. field axioms) while a model is **an** object which satisfies the theory (e.g. a particular field). **It** is an **interesting** question as to whether a theory has a **unique** model. A way of **thinking** is that formal theory hopes to have thinking as a Jnode; or physics' theory hopes to **have** the world as a **model** (rather than the theory has a model which closely approximates the world). Usually, I think, the model precedes a theory. A model can be **thought** of as **the** substructure to a theory; that is, a theory is **an analysis** of **some** model.

For **those** areas which are rigorously **defined**, a model **in** AI has the same sense (for me) that it has **in** logic, except that I would **maintain** the emphasis that a model motivates **formalizing** a theory. **In** most areas of **science**, **the** model really is some domain of real world behavior, and the **game** is to devise **an** approximate model which motivates a theory.

## RICHARD

Well, let **me** put it this way: how, for you, is a model different from a data base?

TOM

Well, to be a model a data base **must contain only** mutually **relevant and coherent** data,

RICHARD

I'm still puzzled as to what you **mean** by model. What properties does this **coherent** data base have that makes it peculiarly a model? **In** particular is it just an **uninterpreted language** or its **interpretation**?

TOM

Its both the **language and** its interpretation. Geometry is a good example of a model in my sense.

RICHARD

What is geometry a model of? It seems to me a model must be a model *of something*. Is it a model of the world?

TOM

**No!**, **not at all**: geometry for **me** is a **model** of my (or anyone's) **computational** structure. The data structure contains declarative **information and computational** procedures (e.g. vector addition).

RICHARD

Let **me put my basic question** another way: clearly a model for you then is **not** just a data base but **includes the action of computing** not just the **linguistic description** of things, its **active**.

TOM

Yes, as **I** said, it includes procedures.

YORICK

**Tom**, this **model, when** you've got it, is **then** a model of the **person's** computational structure not of the real world direct. **Then** words, for you, refer to these models, **and then** the models, by **some** looser relation, refer to the real world?

TOM

**Right**, though I have a model in me of the world, too. But **not necessarily** of the whole world.

RICHARD

So you don't feel any **need** to say *why* this structure is a **model** of a **human's mental** activity -- you **just** say it is -- and the tests are behavioral?

YORICK

And here's a big **difference** in senses of "model" because I think that models, to be models at all, **must** have some **point** by **point correspondence** with what they model -- and Tom **doesn't**.

RICHARD

Right, **Tom** has a behavioral view of **models**.

TOM

No, the' **function** of models is to predict the future. So they must **correspond** in a strong sense to the world, but are not isomorphic with it. The models only mimic a **portion** of behaviour. I **contend** that the **assumption** that we know **only** models, not objects directly, forces us to this view.

YORICK

I disagree strongly, that's a theory that you're talking about, Models **only** predict the **construction** of theories -- see Mary Hesse **on** induction over **scientific** models. Theories predict the consequences of experiments -- that's a basic **difference**, Tom, unless you just **want** to use the two words **interchangeably**.

TOM

Well, we disagree about the meaning of "**model**".

BRUCE

Ok, so we differ about that -- but I **think it needs** stressing here that in spite of this verbal, or labelling, difference, probably all of us want to build the **same** kind of active **computational** objects that **Tom** calls models. All that's at issue is the formal expression of what he calls the "coherence" in the model.

RICHARD

Tom, I **think** both Bruce **and** I are a little surprised by what you seem to be **saying**. Am I **right in thinking** that you believe that the **solution** of several **individual** AI type problems **can add up to a general** solution. I'd like to explore this, since it certainly influences **how** we do (or should do) research.

TOM

AI **encompasses** so many **areas** that there is **no one** model for AI. For each of **many** areas there are models, most of which do **not** yet exist. **Some** of the domains are **quite formal**, e.g. geometry and algebra. In these **domains**, the models are the same as those of logic. In **some** domains, models may exist without **any** theory, i.e. without any analysis. For most domains, there are **no** formal models **now, and** in many areas, we do **not** expect any formal **models**. We cannot really use **the same sense** of model as logicians for these domains, although the **analogy** **springs** from our desire to **represent** these areas as simply **and** compactly as we can. We **really mean then** that a model is a **coherent** body of knowledge about some limited **domain**. In reality, it is just a data structure. The form of the data structure is the **representation** of the **domain**. There is a group of workers devoted to **representing knowledge** without **specifying** what that **knowledge** is. The more meaningful work, to me, is **representing knowledge** about particular domains, e.g. shape of objects,

ARTHUR

No, Tom, I think you're **misrepresenting** that particular group. They are not trying to represent knowledge **in abstract**. They're **much more** concerned, as philosophers have **also** been, with exploring the adequacy of various languages for capturing **epistemological** structures **iii** a large **number of domains**. They are **concerned** with exploring the limits of what can be said: but that's quite different **from** the rather malicious way in which you've **characterised** them. But let that pass.

BRUCE

There is a **confusion** here **between** two uses of the word "**geometry**". One refers to **ur-geometry**, our informal (and mostly **unconscious**) knowledge about **straightness, parallelism and so on** (needed for example by our visual system for perspective, occlusion etc), **and** the other to the semi-formal theory we **learnin** geometry lessons at school. Perhaps you see the **latter as a formalization** of the former, but I don't **think** this is at ail obvious. **In** fact I **don't believe it!**

TOM

At **thispoint**, we want to start asking "model of what?". But we **won't** The **meaning** of 'adequate' is that it predicts a coherent **and** extensive body of measurable **phenomena**, i.e. **relat inns among observables**. It **should** also be adequate in the sense of not having **demonstrable** inadequacies (it should be capable in **principle** to **understand** the **broad range** of behavior). It is **impossible to** prove, only to disprove a theory, **and** the way to go about that is **look** at its structure and test out the **independent predictions**. The sense of **extensive prediction** is that **there** are **predictions** which are independent of any of the data **on** which the theory is based. (One school of physics **maintains** that a theory should **only involve measurable**.) Often models are **taken** because they are **analyzable** (linear economic models) and **not** because they are adequate. This is **frequent in engineering and social science**. There are **models** which are **unanalyzable**, because the relations are too complex for **mathematical** analysis, or because the relations are not well-defined, or because the model is **incomplete**. A **phenomenological** model may describe the results of a coherent set of **measurements**, without **any sense** of **describing unrelated phenomena** at ail. This is a type

of curve fitting, and hasn't any relation to any fundamental structure. It may have utility for experimentation, engineering, or functioning as a biological creature. But phenomenological models are just a form of paraphrasing facts. In many fields, the word theory is used for any trivial explanation: one fact, one "theory". I would instead call this paraphrasing facts, too. There are, in this sense, very few fields with any theories at all. Probably AI has none, although a few extremists might say that predicate calculus with resolution is a theory of reasoning.

BRUCE

Tom, you put too much emphasis on the independence of microtheories, and correspondingly not enough on "knowledge about knowledge" and joining up microtheories (if that is the right way to look at it!). If we can do only a bit of the latter, it is a great help in the former as it tells us about the form of the microtheories and helps us develop them. This sounds a bit like the Richard/Arthur position that if you don't have some coherent overall theory (for them, a logic) then you can't make much progress in the individual areas, but of course I wouldn't want to go that far!

YORICK

But to do this, and create a real theory of language, rather than just a coding scheme, AI must give up another bad habit: the "Queen of the Sciences" thing -- assuming that to deal with language you must express all the information expressible in it -- Tom believes this I think (random thought: how would you keep out the bad theories like Phrenology and Astrology? How about Religion?) I think AI must recognize its task as expressing twentieth-century common-sense knowledge.

TOM

Yorick has said that AI should avoid the pitfall of the "Queen of Sciences". I contend that it should avoid the pitfall of thinking that there is a short cut to intelligence without being the "Queen of Sciences". This does not mean that we must extend the frontiers of chemistry, astronomy, physics, neurophysiology, etc. We must, however, incorporate the common sense knowledge (a vast structure) in all of us, and the many models included there. I would ask, what does understand mean? Given any limited domain, we can tailor the knowledge necessary to eliminate knowledge of some domains, but to include all the functions of human intelligence, we will need all the structures, not necessarily in one system. If the system has a really wide range of intelligent function, then most of these structures will need to be there together.

I conceive of the representations embedded in some sort of deduction system. Clearly that system has considerable importance: it is not clear that the deduction system is uniform over all models or that it is neatly separated as I suggest. We conceive of domain specific representation as primary. We want to go on to say how we conceive it possible to form these models. There seem to be two paradigms: the deduction/induction paradigm which contemplates a static world and draws long chains of conclusions; and the passive observer who watches things happen and draws conclusions, presumably on a statistical basis. The experimental paradigm is the only one I consider at all relevant to building models of the complexity necessary. In this case, the system plays with objects, etc., systematically

varying variables according to independence assumptions, drawing hypotheses which are verified by these simple experiments. In this paradigm, none of the chains of reasoning are very long, and they are immediately verified.

ARTHUR

But what are the implications of this for a theory of meaning?

TOM

From this viewpoint, language is merely a low-quality link of our models to those of others. Meaning in language is a pointing to models which point to other models and eventually to objects. That is, we think that language is totally referential, but referential to models. There is no way we can refer directly to objects, but our intent is that our models really stand for something out there. Reference cannot be infallible: any changes which happen faster than we can perceive can be fooled. If I leave, the orange which sits where I left an orange may not be the same orange. It is not even possible to say that the orange is "similar" to any other, except by the grossest flights of belief. In fact, the payoff seems high in proceeding as though we could, and the risks (with oranges at least) are not large. Although language is notoriously weak in bandwidth and expressing our impressions, it is significant that a large part of our high level models come from language, and it almost seems as though that mode is as important as our perceptual systems.

BRUCE

Hmm. For me language is a very high-quality link. The structures inside people's heads are very complicated, and no two are exactly alike (that's why meeting new people is fun!) so that there is no practical way to integrate my question about so-and-so's tallness into your data-structures. The interface we go through is language and the amazing thing is how much we can conjure up in other people's minds with so few words!

I think we should end our discussion by bringing it down to earth and saying how the ideas we've talked about should affect research programmes. How about a description of a robot system that we can all agree on, and then bringing out the differences?

RICHARD

A brilliant idea, couldn't have thought of a better one myself.

YORICK

I'll second that, old Bruce is really miles ahead of us in the clarity of his thinking!

ARTHUR

But very much behind Wittgenstein, Hintikka, Davidson, Plato, Aristotle, Strawson and Thomas.

RICHARD

Thomas? Which Thomas? Thomas Aquinas?

TOM (interrupting)

Less of this philosophical bullshit, let's hear what Bruce has to say.

BRUCE

Thanks, Tom. You **seem** to be the only (other) **sane one** here!

YORICK

Seriously **now**, could we **end on** a more general **note**? I'd like to hear some final observations from people **on** the following: the last **discussion** about models has shown **me something** I'd rather **not** have **seen** and also answered Bruce's earlier **question** about the **relation** between Phil-marl and Robo-man. I find myself in more or less complete **agreement** with **Tom** about the sorts of active **nondeductive** models we'd like to build or see built. At the same time I disagree **strongly** with him about the *metaphysics* of models and refereiice -- **in that** he like Arthur believes that **meaning** is really **reference**, while I believe that it is some **internal** feature of the whole **language** system.. All that **tends** to suggest that the **metaphysics** are **independent** of the **constructive** activity **and I don't** really **want** to believe that. Further **confirmation** of that **nasty conclusion** is to be found in our **varying metaphysical** **versus** coristructive alliances here. I suspect Bruce, Tom **and I** agree **on** what we'd like to see: **something** non-deductive, active, without **strong** theory -- or at least, Tom, **you** said a short **while** ago that you **didn't** believe in **strong** formal theories -- but I'm not sure you really **want** to be **committed** to that. Whereas Arthur disagrees strongly **and** Richard holds **an intermediate** positioii. But metaphysically at least, Arthur and Torn agree, **while conversely**, Bruce and I agree **in** opposing what **Tom and Arthur** agree **on** (chiefly the **metaphysics** of **reference**). Hence the mutual agreements in the two domains are **conflicting**, thus **supporting** Bruce's thesis of the **independence** of philosophy **and** AI.

ARTHUR

I'm sure your **analysis** of our respective positions is accurate, Yorick, but I don't see that **your conclusion** about the independence of AI and philosophy follows at all **from** it. If I **wanted** to be **harsh**, I **might** simply suggest that all of you other people are simply **inconsistent** in your views **on** metaphysics and activity, while I simply hold a strongly **consistent** view. That's **an** unlikely, -but surely plausible, **conclusion**, isn't it? I must add that I consider holding such a strongly **consistent** view to be important simply **on the** ground.5 that it makes me **much** more vulnerable to **definitive** contradiction, which is **an altogether** good thing from a general scieiitific point of view.

YORICK

**No**, Arthur, there's straightforward **misunderstanding** here. I'm **not** advocating that **independence** (of philosophy and AI -- on the coiitrary *our point* of agreeemerit is that we're **both** advocating **consistency**). Bruce is advocating that independence, **and** I'm raising again

for discussion the possibility that he might be right, **and adducing** as evidence the cross **agreements** we have established. You correctly point out that your metaphysics of **reference** is **consistent** with your taste in **models**, **and** I agree. May I point out that that consistency of yours in **no** way contradicts **my** also being consistent in **my non-referential** metaphysics **and** **my** taste in models. You **and** I can disagree **fundamentally and** still both be consistent. Of course, what **I infer** but **didnt** say, is that it is those who *cross-agree with you and me on models and metaphysics* who are **i nconsisten t.**

TOM

I disagree with Yorick's summary of the putative **agreement among** Bruce, Tom and Yorick. I do feel that a formal theory **independent** of semantic domains is **not** enough; that we **must** have detailed structures for individual semantic domains, and that this knowledge is our **dominant interest**. I am **not** opposed to formal system; **on** the contrary, I favor formal systems. I **infer a statement** that there is an **inconsistency between** the metaphysical position (meaning is **model-referential**) **and** the constructive (build those models as **representations** of knowledge) position. I fail to see the **inconsistency**. Nor do I really see **clearly** what our **differences** are: they do **not seem** so clear-cut to me. **When** I have tried to **explore them**, the differences have largely escaped-like steam, **and I remain** with a feeling that we differ strongly in our estimate of the pragmatic values of formal systems, philosophy, and detailed knowledge of **individual** domains. Further, we differ **in** the **problem** areas in which we want to **make** progress, i.e. formal systems, language **translation**, **perception**. **I sense among** us an **intolerance** that I expect among the most delightful people.