Invited contribution:
Joint Session of Mind Association and Aristotelian Society July 1986
Reply was presented by L.J.Cohen, Oxford.
Both published in
Proceedings of the Aristotelian Society, Supplementary Volume LX, 1986
pages 61--80, 81--97.

WHAT SORTS OF MACHINES CAN UNDERSTAND THE SYMBOLS THEY USE?

Aaron Sloman

Cognitive Studies Programme
University of Sussex
Brighton BN1 9QN England
(Now at University of Birmingham, UK)

Introduction

I am grateful for this opportunity to discuss with philosophers some difficult issues common to philosophy and AI. It is increasingly difficult to keep up with all the relevant literature, and only mutual aid can prevent time-wasting re-invention of wheels and blundering down blind alleys.

My topic is a specialised variant of the old philosophical question ‘could a machine think?’. Some say it is only a matter of time before computer-based artefacts will behave as if they had thoughts and perhaps even feelings, pains or any other occupants of the human mind, conscious or unconscious. I shall not pre-judge this issue. The space of possible computing systems is so vast, and we have explored such a tiny corner, that it would be as rash to pronounce on what we may or may not discover in our future explorations as to predict what might or might not be expressible in print shortly after its invention. Instead I’ll merely try to clarify what we might look for.

Like Searle ([11,12]) I’ll focus on a specific type of thought, namely understanding symbols. Clearly, artefacts like card-sorters, optical character readers, voice-controlled machines, and automatic translators, manipulate symbols. Do they understand the symbols? Some machines behave as if they do, at least in a primitive way. They respond to commands by performing tasks; they print out answers to questions; they paraphrase stories or answer questions about them. We understand the symbols, but do THEY?

Is real understanding missing from simulated understanding just as real wetness is missing from a simulated tornado? Or is a mental process like calculation: if simulated in detail, it is replicated?

If ‘understanding’ denotes some logically private internal state which can only be defined ostensively by pointing inside yourself, then the question whether machines can understand becomes undiscussable, like the question whether the earth is or is not at the same point in absolute space as it was a year ago. Two people ‘pointing’ inside themselves cannot be sure they are talking about the same thing when they ask whether machines, or even other people, have it. Arguments from analogy need a theory which indicates why certain common aspects of the body or brain might be sufficient to produce understanding. And that requires some kind of non-ostensive, functional, analysis of what understanding is, just as questions about identity of locations require locations to be relative to a framework of reference.

In that case, understanding is defined in terms of a collection of capabilities with a certain structure and certain functions. It is not a simple state, and it may be present in different degrees of sophistication. In this functional sense there is a discussable question whether machines could ever understand the symbols they manipulate. This does not imply that there will be a determinate answer.

We’ll see that our ordinary concept of ‘understanding’ denotes a complex cluster of capabilities, and different subsets of these may be exhibited in different people, animals or machines. To ask ‘which are necessary for REAL understanding?’ is to attribute spurious precision to a concept of ordinary language. Hence there is no clear boundary between things that do and things that do not understand symbols.

I shall list ‘prototypical’ characteristics of human use of symbols with understanding, and discuss conditions under which these characteristics might be found in machines. Then instead of answering either ‘YES’ or ‘NO’ to the question whether suitably programmed computers can understand, we note that within the space of possible ‘behaving systems’ there are indefinitely many cases, some sharing more features with human minds, some fewer. The important task is to analyse the nature and the implications of these similarities and differences, and not to argue about which cases existing labels ‘really’ fit.

The space of possible systems is not a continuum. There are many discontinuities that make a difference to functional capabilities. So we are not talking about differences in degree, like differences in speed or memory size, but differences in structure and function, like the difference between having eyes and not having them, or the difference between legs and wheels. There is not just one crucial division. There are very many differences between amoebas and people, no one of which is the ‘essential’ one which makes us conscious or intelligent, just as there is no one ‘essential’ difference between chess and football.

Analysing complex capabilities and distinctions in the space of possible systems can help theoretical biology by presenting a framework for questions about the evolution of behaviour. It can help psychologists by clarifying the nature of the capabilities they are attempting to study. It can help computing science and artificial intelligence by identifying precise new engineering targets for the future.

Philosophers can help by identifying confusions, gaps and errors in the analysis of capabilities we all know about, and extending the analysis to a far wider range of mental concepts. Ordinary philosophical analysis needs to be extended by adopting what Dennett [3] calls the design stance. For example, by analysing possible computational mechanisms instead of just behavioural or phenomenological analyses, we can hope to achieve theories with greater generative power, and therefore greater depth and clarity.

II

What is understanding a language?

I use the word ‘language’ loosely as equivalent to ‘notation’, ‘representational scheme’, ‘symbol system’ etc. Very roughly, a language L is a system of symbols used by some agent U in relation to a world W. ‘System’ implies a generative notation, with compositional semantics. I use ‘agent’ without implying purposiveness. For now I want to leave it open whether the use of symbols presupposes purposiveness in *all* cases, though it obviously does in some. The word ‘use’ may be thought to imply purposiveness, but I intend it to be taken in the sense in which a plant uses oxygen, without having any purpose or intention.

A voice-driven juke-box would relate spoken numbers to a world of records in a rack. The juke box’s limited behavioural repertoire makes it inappropriate to describe it as doing anything more than relating certain symbols to objects. It does not, for instance, relate symbols to properties, relations or states of affairs. Its world W contains a very restricted class of events, and every symbol refers implicitly to that class, merely indicating which object (which record) should partake in the event. The sorts of symbol-users we’ll be interested in will generally be far more sophisticated. They may be able to use symbols in L to refer not only to objects but also to different properties, relations, events, processes, or actions in W. And they can do different things with their symbols.

In simple cases evidence that U uses a symbol S to refer to object O consists of bi-directional causal links. (a) occurrences of S, manipulated by U, may cause U to do something involving O. For instance, finding ‘the big red block’ in an input string may cause U to pick up a certain block. (b) a happening involving O may cause U to do something with the symbol S. Sensors detecting that a certain block is moving might cause U to build a structure containing the string ‘the big red block’.

Symbol manipulations need not be externally detectable. A computing system may do things internally which cannot be inferred from its behaviour, and it may have neither tracing programs, nor access to an output medium capable of displaying the internal detail (see ch 10 of [13]). Unlike behaviourists I am talking about the very kind of internal behaviour which behaviourists try to analyse away.

A full analysis would distinguish different kinds of: (a) symbol media, (b) grammatical rules (c) semantic rules (d) mechanisms for manipulating symbols, (e) symbol users, (f) worlds, and (g) purposes for which symbols might be used. This paper discusses only a subset of this rich array of possibilities.

Symbols are structures that can be stored, compared with other structures, searched for, etc. They may be simple or complex (i.e. composed of parts which are symbols). They may be physical, like marks on a piece of paper, or virtual symbols, i.e. abstract structures in a virtual machine, like 5-D arrays in a computer (See [17]). They may be internal or external. They need not be separable physical objects or events, since a single travelling wave may 'carry' different symbols simultaneously, and a network of active computing nodes may have several patterns distributively superimposed in its current state in holographic fashion. A set of bits may represent one Godel number corresponding to a set of sentences.

Symbols include maps, descriptions, representations, of all kinds, including computer programs, and non-denoting symbols, like parentheses and other syntactic devices. (In fact, anything at all can be used as a symbol.)

The symbols need not be used for external communication. Meaning and understanding are often assumed (e.g. [8]) to be essentially concerned with communication between language users. As argued in [14], this is a mistake, since understanding of an external language is secondary to the use of an internal symbolism for storing information, reasoning, making plans, forming percepts and motives, etc. This is prior in (a) evolutionary terms, (b) in relation to individual learning, and (c) insofar as the use of an external language requires internal computations. Representation is prior to communication.

Objects in the world W may be concrete (e.g. physical objects) or abstract (e.g. numbers, grammatical rules). They may be external, or internal to U. W need not be uniquely decomposable into objects, relations, etc. E.g. a human torso is not uniquely decomposable. Like symbols, the objects may exist in a *virtual* world, embodied in a lower level world, like a virtual machine implemented in a lower level computer. Many programming languages refer to objects in a virtual world, such as lists, arrays, procedures, etc. Similarly social systems form a virtual world embedded in a psychological and physical world.

III

The structure of the concept 'understanding'

A prototypical set of conditions for saying that U uses some collection of symbols as a language L referring to objects in a world W is presented below. Different combinations of conditions define different concepts of 'language', 'meaning', 'understanding', etc. Asking which is the 'RIGHT' concept is pointless.

Some are 'structural' conditions concerned with what mechanisms for understanding do. Some are 'functional' conditions, concerned with what understanding is used for, and how the mechanisms contribute to a larger functional architecture.

We can treat the conditions as a set of axioms implicitly defining 'use of symbols with understanding'. We'll see that events and processes in a computer can constitute a model for a significant subset of the axioms. Moreover, it is not just an abstract model. Unlike simulations of (e.g.) tornadoes, computer models of mental processes can have the same causal relations to the rest of the world as natural mental processes. People outside the model can relate to a machine model as to the real thing (though some may not wish

to). A robot may obey commands, answer questions, teach you things. But a simulated tornado will not make you wet or cold.

We'll see that computers can manipulate internal structures and use them as symbols associated with what Woods, in [21], calls a 'directly accessible' world W consisting of both entities within the machine and more abstract entities like numbers and symbol-patterns. (Cohen [2] also points this out.) Later, we discuss reference to an 'external' world.

IV

Prototypical conditions for U to use L to refer to W

- L is a set containing simple and complex symbols, the latter being composed of the former, in a principled fashion, according to syntactic rules.

- U associates some symbols of L with objects in W, and other symbols with properties, relations, or actions in W.

These condition are satisfied by most computer languages, though machine codes generally have very simple syntax. A computer can associate 'addresses' (usually bit-patterns) with locations in its memory (possibly a virtual memory) and other symbols with their contents and relationships. The symbols cause processes to be directed to or influenced by specific parts of this internal 'world' W. Some of the symbols specify which processes - i.e. they name actions in W.

Various sorts of properties and relations may be symbolised in a machine language, e.g. equality of content of addresses, neighbourhood in the machine, arithmetic relations, etc.

Instructions have imperative meanings because they systematically cause actions to occur. They may have independently variable components, e.g. object, instrument, manner, location, time, etc.

If U is a computer and L its machine code, the semantic relation is causal:

'S refers to O for U' =
    'S makes U's activities relate to or involve O,
    and facts involving O affect U's use of S'

where O may be an object, property, relation or type of action.

- Some objects referred to in world W may be abstract, e.g. numbers.

Computers can use certain symbols to denote numbers because they are manipulated by arithmetical procedures and used as loop counters, address increments, array subscripts etc. (Compare [2].) Computers can count their own operations, or elements of a list that satisfy some test. This has much in common with a young child's understanding of number words

- they are just a sequence of symbols used in certain counting activities ([13] ch.7).

- What a complex symbol S expresses for U depends on its structure, its more primitive components and some set of interpretation rules related to the syntactic rules U uses for L. I.e. L has compositional semantics ([6])

This is true of many computer languages. E.g. what is denoted by a complex arithmetical expression, or a complex instruction, depends on what the parts denote, and how they are put together according to the syntactic rules of the language.

- A distinction can be made between the reference and the sense of symbols, i.e. between what they refer to and how they refer.

A simple example to be found in computers would be the difference between two numerical expressions which necessarily denote the same number, but as the result of different calculations. Similarly, two expressions may access the same internal data but via different routes.

- It is sometimes suggested that real use of a language requires that the mapping between symbols and objects be arbitrary, e.g. unlike ‘clouds mean rain’.

This is partly true of computer languages. However, total arbitrariness would be inconsistent with compositional semantics, and the use of systematic names.

- U can treat the symbols of L as ‘objects’, i.e. can examine them, compare them, change them, etc., though not necessarily consciously.

This applies to computers. Symbolic patterns used to refer can also be referred to, compared, transformed, copied, etc. It is not clear whether other animals can or need to treat their internal symbols as objects. This may be a pre-requisite for some kinds of learning.

- Certain symbols in L express conditionality.

This underlies flexible and creative thinking, planning, or acting. We can distinguish (a) ‘if’ used in conditional imperatives, (b) ‘if’ used as the standard boolean (truth-functional) operator and (c) ‘if’ used in conditional assertions. (c) is not found in the simplest computer languages. (a) and (b) are found in machines.

- By examining W, U can distinguish formulas in L that assert something true from those asserting something false.

Computers typically use symbols for Boolean operations e.g. ‘or’, ‘and’, ‘not’ and two ‘truth-values’. They are taken as truth-values partly because of their role in conditional imperatives. Truth-values can be assigned by examining internal states or arithmetical relations.

- U can detect that stored symbols contain errors and take corrective action.

E.g. programs can attempt to eliminate wrong inferences derived from noisy data, e.g. in vision, and plan-executors can check whether the assumptions underlying a plan are still true. This supports a richer conception of a truth-value than just two arbitrary symbols.

- A complex symbol S with a boolean value may be used for different purposes by U, for instance: questioning (specifying information to be found), instructing (specifying actions), asserting (storing information for future use).

S functions as a primitive question in a conditional instruction where action depends on the answer to the question. In low level machine languages there is not usually the possibility of using the same symbol to express the *content* of an imperative as in "Make S true". I.e. machine codes do not have 'indirect imperatives' with embedded propositions. However, AI planning systems do. (See [Boden 1978] for a survey]). Most computer languages include requests and instructions, but not assertions. However, it is easy to allow programs to record results of computations or externally sensed data, or even results of self-monitoring.

The symbol S may specify the content of an assertion in one context ('store(S)'), a question in another ('if S then...' or 'lookup(S)'), and an instruction in a third ('achieve(S)'). I.e. role is determined by *use* rather than form or content. (This mirrors the distinction between mental states and their contents.)

- U can make *inferences* by deriving new symbols in L from old ones, in order to determine some semantic relation (e.g. proofs preserve truth, refutations demonstrate falsity).

Work in AI has demonstrated mechanisms for doing this, albeit in a restricted and mostly uncreative fashion so far.

- L need not be a fixed, static, system: it may be extendable, to cope with expanding requirements.

Many computer languages are extendable. Adaptive dialogue systems are beginning to show how a machine may extend its own language according to need. But deep concept formation is still some way off. It is not clear which animals can and which cannot extend their internal languages. Without this, certain other forms of learning may be impossible.

- U may use symbols of L to formulate goals, purposes, or intentions; or to represent hypothetical possibilities for purposes of planning or prediction.

Simple versions of this sort of thing are AI planning systems. Only a system whose functional architecture supports distinctions between beliefs, desires, plans, suppositions, etc., can assign meanings in the way that we do. Merely storing information, and deriving consequences, or executing instructions, leaves out a major component of human understanding, i.e. that what we understand *matters* to us. For information to matter to a machine it must have its own desires, preferences, likes, dislikes, etc. This presupposes that

there are modules whose function is to create or modify goals - motive generators. Full flexibility requires motive-generator generators. Deciding and planning require motive comparators and motive-comparator-generators. This is spelled out a little more in [15]. Motives generated internally over many years, refute the claim that a machine can exhibit only desires of the programmer or user. Such a machine would use symbols in L for *its* purposes.

This is an important boundary in the space of possible behaving systems. Without this structure a machine might understand well enough to be a slavish servant, but could not be entrusted with tasks requiring creativity and drive, like managing a large company or minding children.

- L may be used for communication between individuals. This adds new requirements ([21]), which I shall not discuss, since representation is prior to communication.

All the conditions so far listed for U to use a language L in relation to a world W are consistent with U being a computer. Several do not even require AI programs, since modern computers are built able to use symbols to refer to a world W containing numbers, locations in memory, the patterns of symbols found in those locations, properties and relations of such patterns, and actions that change W.

V

Does the computer really understand?

Searle's claim that computers appear to understand only because people interpret the symbols, i.e. the process has only 'derivative' intentionality, ignores the fact that a substantial portion of the structure of the concept of 'using a symbol with a meaning' is exemplified even without AI programs. Associations between program elements and things in the computer's world define a primitive type of meaning that *the computer itself* attaches to symbols. Its use of the symbols has features analogous to simpler cases of human understanding, and quite unmatched by juke boxes. So it does not interpret symbols merely derivatively: the causal relations justify our using intentional descriptions, without anthropomorphism. To simulate or replicate human types of intentionality, including beliefs, desires, plans, fears, attention and self-consciousness, requires the embedding of individual mental processes in a suitable network of co-operative processes with intricate divisions of functions.

In short, though *structural* requirements for at least the simplest sorts of understanding are relatively easy to achieve, *functional* requirements are harder. We know how to make mechanisms capable of producing intentional states. However, to be intentional processes like human mental processes, the symbol-manipulations must themselves have additional causal powers: the power to affect beliefs, desires, plans, and the actions they produce. This requires connections with additional procedures and data-bases concerned with the use of symbols in a manner characteristic of beliefs, desires, plans, etc. All this is possible even if W is a purely internal world, like the world of a dedicated, enthusiastic mathematician.

VI

Reference to inaccessible objects

Machines can refer to their own internal states, to numbers, and to symbolic patterns, i.e. what Woods [21] calls a 'completely accessible' world because semantic links between symbols and things in this world are directly derived from simple causal links and the way symbols are used. In order to be useful as robots, or friends, machines will need to refer to external objects, events, locations, etc. The problem of *external* semantic linkage is harder to deal with.

How can a system use symbols to describe objects, properties, and relationships in a domain to which it has no direct access, and only incomplete evidence, so that it can never completely verify or falsify statements about the domain (like unobservables in physics)? Some external reference uses external causal links, such as sensors and motors. But direct links are often impossible, e.g. referring to events remote in space and time, or even to hypothetical objects in hypothetical situations. What alternative types of semantic link might there be?

A key idea is that implicit, partial, definitions (e.g. in the form of an axiom system) enable new undefined concepts to be added to a language. (Compare [1]) on 'meaning postulates' and [21] on 'abstract procedures'.) A collection of axioms for Euclidean geometry, in the context of logical inference procedures, can partially and implicitly define predicates like 'line', 'point', 'intersects', etc. The axioms constrain the set of permissible models. Similarly, a congenitally blind person may attach meanings to colour words not too different from those of a sighted person, because much of the meaning resides in rich interconnections with concepts shared by both, such as 'surface', 'edge', 'pattern', 'stripe'.

A Tarskian semantic theory does not, in general, allow meanings to be fully determinate, since it will always be possible (except in very simple cases) to add further axioms constraining the possible models, and adding precision to the meanings of the terms. It is also generally possible to add axioms postulating additional entities and new relations between those entities and the previous ones, just as science advances partly by postulating new sorts of entities: atoms, genes, etc.

Combining our previously discussed internal causal links with Tarskian semantics, allows symbol-users to refer to their own internal states and also to very general possible states of possible worlds. This would permit mathematical thoughts and inventing possible physical universes and engaging in hypothetical reasoning about their inhabitants, properties, etc.

Are external causal connections required for thoughts about particular objects in the environment? (Compare Woods, McDermott [16])

VII

Causal links are required for reference to actual particulars

No matter how many new symbols and axioms are added, Tarskian semantics will not of itself force the symbols to refer to any particular bit of reality rather any other actual or possible bit of reality which has a similar structure and a similar network of relationships.

So the meanings defined simply by a set of axioms will always be totally universal, unless some of the symbols have a different sort of meaning, which attaches them to some individual portion of reality, for instance symbols whose causal connections enable a machine to refer to its own innards, as described above.

Even without links through sensors and motors, an intelligent system might have symbols for a number of general relationships defined axiomatically, which could be used to express thoughts about how portions of the internal world are related to inaccessible objects. Examples of such relations are ‘causes’, ‘before’, ‘inside’, ‘beyond’. How exactly ‘cause’ might be defined axiomatically is an old and unsolved problem. A sophisticated reasoning system might use the meta-level notion of a type of relationship whose detailed definition is not known, to build descriptions of relationships (of unknown types) between accessible objects and others (possibly of unknown types). Such a thinker might think of its own internal states as embedded in a larger structure, and start speculating about the properties of that structure, which it could refer to as: ‘this world’.

Symbols causally linked to input and output transducers (sensors and motors) would have the ability to anchor reference to external particulars. Another example would be the use of demonstratives like ‘here’ and ‘now’ (and implicit use of such things in tensed verbs), which are linked to portions of space and time merely through the spatio-temporal nature of the system using them. (Compare Evans [4].)

Attachment to specific portions of reality can be inherited by axiomatically defined terms, provided the axioms link them to other terms which have a more direct link. This does not imply that the external descriptors are explicitly definable in terms of symbols describing ‘sense-data’ as phenomenalists have supposed. (For more on this see [13] chapter 9.)

Moreover, the inherent indeterminacy of Tarskian meanings explained above can never be totally removed by links to symbols with more direct semantics. At best the indeterminacy will be partially reduced. For example, links between the concept ‘electron’ and what we can observe in a range of experiments leave it open for the concept to be further specified in the future by theoretical and empirical discoveries concerning the internal nature of electrons and their causal powers.

VIII

Loop-closing semantics for non-propositional symbols

I don’t really believe that birds, baboons or babies use logic with Tarskian semantics to enable them to perceive and act on things in the world. Yet there is no doubt that many animals have rich mental lives including thoughts of external objects. Might something other than logical and propositional representations explain this?

A generalisation of Tarskian semantics may be more generally applicable to intelligent systems. There is no reason to suppose that all internal representations must be propositional. There are good reasons for using a variety of forms of representations, including analogical representations such as diagrams, maps, ordered lists, etc. (See [17]).

We can define a non-Tarskian model for the internal representations which play a role in percepts, beliefs plans, etc., namely an external environment which can coherently close the feedback loops. This notion of coherent causal closure will be relative to the system's ability to have precise and detailed goals and beliefs. How specific the mapping is between internal representations and external structures will depend on how rich and varied is the range of percepts, goals and action strategies the system can cope with.

Like Tarskian semantics, 'loop-closing semantics' leaves meanings indeterminate. For any level of specification at which a loop-closing model can be found, there will be many consistent extensions to lower-levels of causal structure (in the way that modern physics extends the environment known to our ancestors), which remain adequate models in this sense. Even for a given level of description the internal representations may be more or less specific: for instance there will generally be infinitely many possible hidden extensions to visible portions of objects consistent with what you know about the world. Your friend may have warts under his shirt.

The notion of loop-closing semantics presupposes a computational architecture rich enough to support distictions between different sorts of internal causal roles of symbols, in particular distinctions between (a) established beliefs (including percepts), (b) hypotheses awaiting confirmation, (c) goals, and (d) plans and instructions. It is far from obvious what sort of design can support such role distinctions, and the consequential loop-closing model theory.

Some causal link is required if symbols are to refer to particular physical objects, like the Tower of London, or physical properties found in our world, such as magnetism. Without causal connections with the environment a thinker could only think (existentially quantified) thoughts about an abstract possible world, or very abstract and general thoughts about this world.

External links differ in kind. Besides visual, tactile, and other sensory links it is possible to have communication with other agents via a keyboard or other devices. I believe these are also capable of pinning down reference. Causal links can be more or less direct, and can convey more or less rich information. Communication via another agent is indirect, and generally provides limited but abstract and general information, but it is still a causal link, like fossil records.

So, using symbols to refer to an external world does not require that the world actually be directly sensed and acted on by the specific symbol-user.

IX

Extending 'mentalese': concept learning

A language may be extended by the addition of new axioms and procedures, partially and implicitly defining some new primitive symbols, and modifying the meanings of old ones. The history of concepts of science and mathematics shows that not all newly-acquired concepts need be *translatable* into one's previous symbolism.

After such learning, there is no clear functional distinction between the original concepts and the accreted language: we can memorise facts, formulas and instructions in English, instead of always having to translate into 'mentalese'. Hence, contrary to Fodor [5], different humans (or machines) may use different 'mentalese' even if they all started

off the same.

X

The essential incompleteness of semantics

We have seen that both Tarskian and loop-closing semantics leave symbols with partially indeterminate meanings. Causal links, like added axioms, reduce, but do not remove, the indeterminacy. This incompleteness is evident in theoretical concepts of science, but can also be demonstrated in ordinary concepts.

In a sufficiently complex thinking system, even the language used for describing its own *internal* state will have this kind of indeterminateness and incompleteness, because of the problems of internal access sketched in chapter 10 of [13].

XI

Can a computer distinguish ‘true’ and ‘false’?

It is not clear how to distinguish a ‘true’ from a ‘false’ boolean value, since formally they are symmetrical. The manual may say that 1 stands for ‘true’, but formally 1 could equally be interpreted as ‘false’, 0 as ‘true’, ‘and’ as ‘or’, ‘or’ as ‘and’ etc.. Could there be an asymmetry in the use of the symbol for ‘true’ and the symbol for ‘false’?

One source of asymmetry lies in mechanisms that check assertions, instead of always blindly assuming them correct: an elementary form of self-consciousness. ‘True’ might label a tendency to survive thorough checking. But the connection is not simple, for the result of checking may be wrong.

A ‘redundancy convention’ could produce asymmetry. Instead of using explicit booleans, adopt a convention that one of the boolean indicators is redundant: it is signified merely by the presence of a formula in an information store or a communication. Given negation, ‘true’ and ‘false’ then both become redundant labels.

A deeper asymmetry lies in connections between beliefs and autonomous motives. True beliefs are those which (generally) enable desires to be satisfied by rational planning. Again the connection is not simple, for a true belief can lead to a disastrous plan.

XII

Can understanding be truly duplicated, or only simulated?

Many readers will object to the suggestion that if certain formal conditions are satisfied by the processes in a machine, then it understands. This has been called the ‘Strong AI’ thesis. A common way of arguing against it is to describe a process which conforms to the allegedly sufficient conditions yet clearly does not involve understanding.

One supposed counter-example is a person who does not understand Chinese taking the place of a computer running a program allegedly capable of producing such understanding. Searle [11,12] claims it would not if he were the person. Another type of example might be a subset of the atoms in a giant storm cloud, or some other randomly moving agglomeration - in principle some subset might happen to form a pattern which could be mapped onto the execution of a program. This would not mean that a storm-cloud

had mental states. Another example might be a random number generator which happened to produce a succession of Godel numbers representing states of a machine following the program.

Full discussion of these objections would require analysis of different ways in which a program may relate to processes which 'instantiate' it. Random connections clearly do not have the *reliability* required for a process which plays the role of understanding within an intelligent system. Though it is not so obvious, the same could be said of a process in which John Searle acts as a computer. The lack of reliability would be due to the potential for Searle's motives, beliefs, distractions, tiredness, etc. to interfere with the running of the program. Thus the process would not satisfy the same set of counterfactual conditional descriptions as the process in a fully integrated intelligent system.

A more complete discussion would show how certain sorts of local unreliability may be required, to allow more global processes to interrupt, modify, re-direct, or abort sub-processes if they do not conform to global requirements of the system. Thus local unreliability or unpredictability may enhance global coherence and reliability.

This leads to the conclusion that not every process which happens to have the right formal properties would constitute understanding (or any other mental state). The underlying mechanisms and the relationships to other parts of the system must have the right causal properties. There is nothing to prevent a computer having those properties, as far as I know. But the alleged refutations of the Strong AI thesis involve systems which don't have the right properties. So they are not refutations after all (compare [18]).

If machines are to have mental states and processes of their own, they must have mechanisms with the right dispositional qualities. For example, merely having some kind of giant lookup table which enables an appropriate response to be produced in a very large set of possible situations would not be adequate. Ordinary understanding of a language involves having a capability with infinite generative power, not achievable by a finite condition-action table, even if the table was large enough to survive a lifetime of testing. Understanding involves having dispositions or capabilities which go beyond the behaviour actually produced. (Compare Cohen's distinction beween 'simulated parotting' and 'simulated understanding' [2].)

None of this proves the Strong AI thesis correct, of course. But it shows that setting up the right causal conditions for understanding (or other mental states) is not a trivial matter. Refutations of Strong AI must address themselves to systems where the reliability conditions are satisfied, not just the *formal* conditions.

XIII

Conclusion

A 'design stance' helps to clarify the question whether machines themselves can understand symbols in a non-derivative way. It is not enough that machines appear from the outside to mimic human understanding: there must be a *reliable* basis for assuming that they can display understanding in an open-ended range of situations, not all anticipated by the programmer. I have briefly described structural and functional design requirements for this, and argued that even the simplest computers use symbols in such a manner that the machines themselves associate meanings of a primitive sort with them.

I have shown that a computer may use symbols to refer to its own internal states and to abstract objects; and indicated how it might refer to a world to which it has only limited access, relying on the use of axiom-systems or perception-action loops to constrain possible interpretations. These constraints leave meanings partly indeterminate and indefinitely extendable. Causal links reduce but do not remove indeterminacy.

The full range of meaningful uses of symbols by human beings requires a type of architectural complexity not yet achieved in AI systems.

There is a complex set of prototypical conditions for understanding, different subsets of which may be exemplified in different animals or machines, yielding a large space of possible systems which we are only just beginning to explore. Our ordinary labels are not suited to drawing a definite global boundary within such a space. At best we can analyse the implications of many different boundaries, all very important. This requires a long term multi-disciplinary exploration.

Acknowledgements

The author has a fellowship from the GEC Research Laboratories, and has benefitted from discussions with members of and visitors to the Cognitive Studies Programme at Sussex University, especially Margaret Boden, Steve Torrance, and Bill Woods.

BIBLIOGRAPHY

[1] Carnap, R., *Meaning and Necessity* Phoenix Books 1956.

[2] Cohen, L.J., ‘Semantics and the computer metaphor’ in R. Barcan Marcus, G.Dorn, P. Weingartner (eds) *Logic Methodology and Philosophy of Science VII,* Amsterdam: North-Holland, forthcoming.

[3] Dennett, D.C., *Brainstorms,* Harvester Press 1978.

[4] Evans, Gareth, *The Varieties of Reference,* Oxford University Press, 1982.

[5] Fodor, J.A., *The Language of Thought* Harvester Press 1976.

[6] Frege, G., *Translations from the philosophical writings,* ed. P. Geach and M. Black. Blackwell, 1960.

[7] Hempel, C.G, ‘The Empiricist Criterion of Meaning’ in A.J. Ayer (Ed.) Logical Positivism, The Free Press, 1959. Originally in *Revue Int. de Philosophie,_Vol.4.* 1950.

[8] Lyons, John, *Semantics* Cambridge University Press. 1977.

[9] Pap, A., *An Introduction to the Philosophy of Science* Eyre and Spottiswoode (Chapters 2-3). 1963.

[10] Quine, W.V.O., ‘Two Dogmas of Empiricism’ in *From a Logical point of view* 1953.

[11] Searle, J.R., ‘Minds, Brains, and Programs’, with commentaries by other authors and Searle’s reply, in *The Behavioural and Brain Sciences* Vol 3 no 3, 417-457, 1980.

[12] Searle, J.R., *Minds Brains and Science,* Reith Lectures, BBC publications, 1984

[13] Sloman, A., *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind,* Harvester Press and The Humanities Press, 1978.

[14] Sloman, A., ‘The primacy of non-communicative language’, in *The analysis of Meaning: Informatics 5,* Proceedings ASLIB/BCS conference Oxford, March 1979, Eds: M.MacCafferty and K.Gray, Published by Aslib.

[15] Sloman, A. and M. Croucher, ‘Why robots will have emotions’ in *Proc. IJCAI*

Vancouver 1981.

[16] Sloman, A., D. McDermott, W.A. Woods 'Panel Discussion: Under What conditions can a machine attribute meaning to symbols' *Proc 8th International Joint Conference on AI,* Karlsruhe, 1983.

[17] Sloman, A., 'Why we need many knowledge representation formalisms', in *Research and Development in Expert Systems,* ed M. Bramer, Cambridge University Press, 1985.

[18] Sloman, A., 'Strong strong and weak strong AI', *AISB Quarterly,* 1985.

[19] Sloman, A., 'Did Searle attach strong strong or weak strong AI?', in A. Cohn and R. Thomas (eds) *Proceedings AISB Conference, 1985* Forthcoming.

[20] Strawson, P. F., *Individuals: An Essay in Descriptive Metaphysics,* Methuen. 1959.

[21] Woods, W.A., 'Procedural semantics as a theory of meaning', in *Elements of discourse understanding* Ed. A. Joshi, B. Webber, I. Sag, Cambridge University Press, 1981.