PANEL DISCUSSION
UNDER WHAT CONDITIONS CAN A MACHINE ATTRIBUTE MEANINGS TO SYMBOLS

Aaron SLoman
(Panel Chairman)
University of Sussex

It is easy to produce arguments purporting to show that computers can never be conscious, have purposes, understand anything. One standard response is to claim that since computers can manipulate symbols they have the potential to support any kind of mental state or process, if appropriately programed. The task is then to investigate what sort of programming will transform a symbol user into something with mental states. Much work in AI is concerned with specific subgoals of this investigation, taking it for granted that machines can manipulate symbols with meanings.

but can this be taken for granted? Certainly computers construct and transform patterns. We can interpret them as meaningful, but it does not follow that the computer does, any more than a filing cabinet interprets documents. If interpreting symbols already presuppose being conscious, having beliefs and goals, etc., then the ability to assign meanings to symbols cannot explain these other abilities.

One popular reply is to postulate sensors and motors. No mind or computational process can *refer* to particular events or objects in the physical world without being causally embedded in the world. Perhaps even reference to properties and relations (universals) requires such causal embedding. So if we find suitable causal relationships between internal processes and external events, then this would seem adequate to justify the claim that the internal processes use symbols with a meaning.

but is this necessary for every form of mentality, every meaningful use of symbols? Can't we imagine someone totally paralysed, blind, deaf, anaesthetised, yet conscious, thinking, even experiencing hallucinatory physical sensations - or at least solving problems about number theory? Or do we only <u>think</u> we can imagine this? People often assemble meaningful words into deceptively incoherent phrases, like "the view of the universe from outside time".

1 claim that we should accept the anti-behaviourist intuition that the meaningful use of symbols can be purely internal. This is not just a semantic quibble about the meaning of "meaning". Rather, the point is that different sorts of <u>internal</u> computational architecture have profound implications for different sorts of mentality, independently of external relations. It is significant that computers have been designed in such a way as to support talk of machine-languages, instructions, addresses, conditionals, logical operations, arithmetic operations, and the like. The use of such terms in computer science does not depend on external relations. It is justified by the structure of <u>internal</u> processing. For instance, a computer can use numbers to count internal operations, or to count internal symbols, just as we use them externally.

The issues are surprisingly tricky. Given an uninterpreted computer you may be able to work out which symbols the machine interprets as booleans. but could you tell which one was taken as "true" and which as "false"? Can the distinction can only be made only in terms of a computational architecture rich enough to support a division between beliefs and goals? We don't notice the need for this normally because machines come with manuals which beg the question by telling us which operation is "and", which "or", etc.

We should explore the ability of different sorts of computational architectures to support such notions as reference, instruction, goal, description, truth, plan, desire etc. In particular, we need to understand the conditions under which a machine could have beliefs and goals relating only to <u>internal</u> states and processes, such as the belief that two lists have the same contents, or a goal like counting the number of even integers in an internal list. (The differences between beliefs and goals lie only in their roles in processing.) The relevant architecture need not be directly implemented in a physical machine. It could be a virtual machine implemented in a lower level machine with a simpler architecture. All this contradicts a philosophy which stresses the social, communicative, role of meaning.

bo ordinary digital computers, suitably programmed, seem capable of supporting quite complex notions of meaning and mentality independently of external connections. Paradoxically, it is far less obvious how brains can support the self-referential abilities which enable ordinary computers to be described in mentalistic language. Do brains intrinsically require <u>external</u> connections for their use of symbols with meanings?

I have deliberately invited a biased panel to discuss these issues. When one side of a debate is so easy to argue for, we may hope to profit most by listening to the other side.