New Ideas Psychol. Vol. 1, No. 1, pp. 41-50, 1983
Printed in Great Britain

0732-118X/83/010041-10$03.00/0
Pergamon Press Ltd

COMMENTARY

AARON SLOMAN
Cognitive Studies Programme, University of Sussex, Brighton BN1 9QN, U.K.

INTRODUCTION

Having discussed these issues with the author over many years, I was not surprised to find myself agreeing with nearly everything in the paper, and admiring the clarity and elegance of its presentation. All I can offer by way of commentary, therefore, is a collection of minor quibbles, some reformulations to help readers for whom the computational approach is very new, and a few extensions of the discussion.

WHAT IS ARTIFICIAL INTELLIGENCE?

I'll start with a few explanatory comments on the nature of A.I., to supplement the section of the paper "A.I. as the Study of Representation". Cognitive Science has three main classes of goals (a) *theoretical* (the study of possible minds, possible forms of representation and computation), (b) *empirical* (the study of actual minds and mental abilities of humans and other animals), (c) *practical* (the attempt to help individuals and society by alleviating problems (i.e. learning problems, mental disorders) and designing new useful intelligent machines).

Activities pursuing these three goals are most fruitful when the goals are interlinked, providing opportunities for feedback between theoretical, empirical and applied work. Artificial Intelligence is a subdiscipline of Cognitive Science which straddles the theoretical approach (studying general properties of possible computational systems) and applications (designing new systems to help in education, industry, commerce, medicine, entertainment). Its empirical content is mostly based not on specialised research, but on common knowledge of many of the things people can do — such as using and understanding language, seeing things, making plans, solving problems, playing games. This knowledge of what people can do sets design goals for both the theoretical and the applied work. In particular, an important aspect of A.I. research is task analysis: given that people can perform a certain task, what are the computational resources required, and what are the trade-offs between different representations and processing strategies? This sort of analysis is relevant to the study of other animals insofar as many human abilities are shared with other animals.

The mere observation and classification of forms of behaviour is likely to be shallow unless informed by a theoretical understanding of the types of internal representations and processes required for coping with the environment. So, as implied by Boden, a training in A.I. should help ethologists (and of course

Commentary on M. A. Boden (1983) Artificial intelligence and animal psychology. Vol. 1, pp. 11-33.

psychologists) to do more profound research. Equally, however, empirical studies of the variety and details of forms of animal behaviour should enrich the theoretical study of possible intelligent systems within A.I.

MAPPING TERRAIN vs DRAWING BOUNDARIES

A potentially important omission from Boden's paper concerns the structure of the space of possible minds. In particular, it is a mistake to think in terms of a continuous spectrum from the simplest organisms through the more intelligent organisms, to man. It is a mistake because there are many important discontinuities in the space of possible systems.

For instance there is a jump from computer operating systems which allow one program to run at a time to those which allow more than one. There is a jump from systems which can represent and manipulate only quantitative data to those which can manipulate other symbolic structures such as sentences, pictures, or plans. There is a jump from systems which can only react to incoming information, to those which can store the information for future use. There is a jump between systems controlled by Condition-Action rules, and those which instead of allowing conditions to trigger actions directly, allow them to set up goals, leaving it to a separate process to select the best action to achieve the goal. There is a jump from systems which inevitably attempt to achieve goals once created to systems which can compare and evaluate goals and decide to reject some. There is a jump between single minded systems which pursue plans to the bitter end and those which allow interruptions and suspension or abandonment in the light of new information. Besides these global discontinuities, there are many more detailed discontinuities between types of computational architecture, types of internal languages, types of algorithm, etc. Boden mentions some of the discontinuities but doesn't draw the general conclusion about the structure of the space of possible systems.

Not only is it a fallacy to think of the space as a continuous spectrum. It is also a fallacy to assume that there is any single major discontinuity. So some old questions are quite misguided, for instance the question: "Where should we draw the boundary between animals with minds and those without?" Searching for a single, major, discontinuity may prevent us appreciating the importance of a myriad of minor discontinuities with cumulative effects. My colleague, Jon Cunningham, summed this up usefully with the slogan "Mapping the terrain is more important than drawing a boundary".

CONSCIOUSNESS

Boden claims in her Introduction and penultimate paragraph that the most puzzling feature of consciousness remains unresolved. However, this may be based on the "single discontinuity" fallacy sketched above. Because our language uses noun phrases like "consciousness", "phenomenal experience", "self awareness" and the like, we are tempted to think that they refer to entities or properties which are definitely there or not there, on the model of other noun phrases like "electric current", "the horse power of this engine", "influenza". Introspection supports the illusion. If, instead, we assume that every aspect of human consciousness is analysable in terms of a multitude of interacting processes of different kinds, and allow that different combinations of these sub-processes are possible, we can begin to understand the relationships between many different forms of consciousness in animals and machines. The residual question which we may feel tempted to ask "Does that thing really have this experience?" (said "pointing" inwardly) may prove to be as misguided as the pre-Einsteinian question "Does this event occur in the same place in space as that one?" Just because pointing can identify a location at a time, it does not follow that it makes sense to talk of identity of locations across time, except relative to a framework of objects and events (the same place in the room, in the solar system, in the galaxy – but not just in space).

So we can ask if two organisms have the same experience insofar as we are asking how their current state relates to a complex of earlier and later states and abilities. But if we think different acts of introspective pointing can directly raise meaningful questions about absolute identity of mental states this may just be an illusion. Nevertheless it is to be expected that an intelligent robot, with the ability to monitor its own internal processes to some extent, will be driven into this trap just as we are. It will also be driven (like Newton) to the reification of spatial locations. In both cases this may be because, for non-philosophical, non-scientific purposes, i.e. the ordinary purposes of day to day living, it is most useful to treat spatial locations and mental events in the same sort of ways as other kinds of objects. The same representational systems and inference rules can then be employed in making plans, forming generalisations, etc. (Perhaps a super-intelligent machine will one day be designed which is not tempted by these fallacies.)

I don't claim to have demonstrated that this "relativistic" analogy solves the problems of consciousness. The detailed arguments are yet to be worked out.

But I hope it is clear that the approach is not behaviourist. States of consciousness are not reduced to behavioural dispositions. In saying that having an experience is a state related to other states, processes and abilities, I am talking about other internal, i.e. computational, states etc.

The problems are complicated by the fact that when all the factual and logical problems are resolved, there remains an ethical problem. The questions "Is this conscious? Does it really have experiences? Does it feel pain?" are not mere academic enquiries. In our system of concepts they are inextricably tied up with the practical, ethical questions "How should I treat this? Should I take its interests into account in my deliberations?" People with a strong tendency to disagree on the ethical issues will translate this into disagreement on the question whether machines and other animals "really" have consciousness. The ethical disagreement may be real even when the factual disagreement is illusory: a matter of drawing arbitrary boundaries.

Incidentally, the "single discontinuity" fallacy is to be found as much in debates about whether animals use language as in debates about the nature of mind and consciousness.

SYMBOLIC AND NON-SYMBOLIC COMPUTATION?

Another quibble. While discussing hoverflies, in the section "A.I. as the Study of Representation", Boden suggests that there is a distinction between symbolic and non-symbolic computation, the latter being based, for example on a simply specifiable rule which could be "hard-wired" into the flies' brains. I believe she is mixing up several distinctions in this paragraph: (a) the distinction between hardware and software (which may depend on one's level of description), (b) a distinction between innate and learned rules, (c) a distinction between rigid unconditional rules and flexible rules allowing different kinds of contexts to enter into decisions.

Unlike Professor Boden and some other authorities, I can see no use for a distinction between symbolic and non-symbolic computation. For how else should we define computation if not as the manipulation of symbols, whether hard-wired or not, whether innate or not, whether quantitative or structural? What is important is that her examples draw attention to some of the discontinuities in the enormously varied space of possible systems. I would have liked her to bring out more clearly the distinction between a system whose environment is represented simply as an array of values for a fixed set of variables, and one whose environment is hierarchically structured in such a way that a grammar for an indefinitely large set of different descriptions is required. The two sorts of representations require quite different processes of learning, recognition, and use in guiding actions.

The "array of values" representation might be useful for an organism inhabiting some sort of soup with varying temperatures, amounts of light, and densities of different sorts of chemicals. The total state of the system could be represented as a set of measures, and the goals could be represented as desired intervals, or values, for these measures. All actions would merely change some of the measures.

The fact that for many animals, including human beings, the world does not have this simple structure has profound implications for the design of explanations of behaviour. For instance, instead of a fixed number of measurements, it may be necessary for an organism to be able to construct and manipulate arbitrarily complex hierarchically organised representations, or even more complex non-hierarchical networks. The kinds of mathematics developed mainly for Physics are therefore not suitable for representing internal processes in which such structures are manipulated. Catastrophe theory, designed to cope with discontinuity within the framework of an array of continuous variables is also inappropriate, since it does not provide a formalism for representing structural change where the numbers of components in a complex whole, and relations between them can change.

Instead we seem to need the formalisms and mechanisms developed in Computing Science and formal Linguistics. For instance, the parsing processes of a compiler may have much in common with many perceptual processes, in which a mass of information has to be segmented into parts and their relationships recorded. My impression is that many students of animal and human behaviour still think in terms of much more simplistic models derived from the physical sciences.

Of course, between the computationally most sophisticated systems, and the simplest organisms which represent their world in a set of measurements (if there are such organisms), there are many and varied intermediate cases. A related point is that there may be no sharp boundary between processes of computation, and mere processes in which things change. The space of possible systems has yet to be mapped out and a good taxonomy developed.

ATTRIBUTING EXPLICIT GOALS TO ANIMALS

Boden's discussion of doves suggests that they are not driven by explicit goals. However, even if they don't have the explicit goal of rearing healthy chicks, it does not follow that no other goals are ever explicitly represented. (Boden doesn't say it follows. Neither does she point out that other goals might be explicit.) She does not say exactly what she means by "explicit goal". I take it this is something like a representation which (a) can be compared with a representation of the current situation, and (b) if there is a mismatch has the ability to cause planning and action to occur to produce a match. (This ability need not be realised if other goals, or other mechanisms interfere.) The notions of "representation", "match" and "mismatch" all require further analysis.

It seems quite plausible that birds do not have anything like the goal of rearing healthy chicks, for that would require them to represent something like "my chicks have reached adulthood and are healthy", which requires a grasp of very abstract concepts. However, the representation of more concrete short term goals may be within their grasp. The enormously varied actions involved in building nests, hunting for food and rearing chicks seems to require that detailed behaviour be driven by explicit goals (e.g. something like: "catch that worm", "go to nest", "sit on eggs", "deposit food in chick's beak", "hop to that branch"). At this stage it is not at all obvious exactly what the goals are, nor what sort of representation language might be used. For instance, it is unlikely that birds make use of categories which would translate into English words like "worm", "nest", "beak", etc. Their concepts may be much more structural, and less functional and theory laden. It is hard to see how a collection of Condition-Action rules, where the conditions refer only to current states, with no representation of states to be achieved (or prevented), i.e. without explicit goals, could, for example, produce effective nest building behaviour in such varied conditions. Such rules might suffice for the higher-level control, and in particular for the creation of new goals. But at the level of matching detailed physical movements to the structure of the environment and the tasks to be achieved, some sort of run-time, goal-directed, plan synthesis seems to be required.

Any form of behaviour can be represented in Condition-Action rules, provided there are enough rules. What seems impossible is that a bird's brain should somehow have enough rules to cope with all the different spatio-temporal configurations which can occur, without ever having to derive plans from a representation of the situation and the current goal. But I cannot prove that it is impossible. A good refutation of my claim would be a working model of a bird driven by a set of Condition-Action rules of a type which might be either genetically programmed or somehow learned in a bird's life time, without being derived from explicit goals.

Even where animals are driven by explicit goals, it is impossible to infer with certainty what the goals are, or how they are represented, merely on the basis of behaviour. At present we don't even have a good theory of what the alternatives are. All we have are a few tentative proposals in terms of formalisms analogous to predicate logic or semantic nets. How might the goal of being on a branch close to the entrance to a nest be represented? Anything like a picture of the situation would be too view-point sensitive. Can a bird handle a more abstract (Fregean) formalism where relationships are named, as opposed to being depicted? If it constructs a non-pictorial three dimensional analogical representation how does it relate this to the currently perceived view, and to possible movements? A sparrow hopping from a twig to another one close by with a totally different orientation clearly must have a representation which is very precise, since without precision it would not land with its feet in the right place.

And if it has, and makes use of such a detailed representation, whatever it is, is there any reason to deny that it is conscious of the twig?

The discussion of the section "Cognitive Ethology and Computational Concepts" may lead some readers to infer that only the behavour of higher mammals, including apes, suggests underlying planning abilities of the kinds currently studied in A.I. My own impression, from informal observation of a wide range of animals, including some insects, suggests to me that they all possess complex cognitive skills in many ways more sophisticated than anything we currently know how to program. The bee's ability to manipulate its flight so as to land on a flower appropriately is perhaps an example, as is its ability (mentioned by Boden) to store information about routes and communicate this to others through its "dance" on return to the hive. Nesting birds which, although they may not be able to vary their higher level goal structures, nevertheless display in their detailed behaviour considerable flexibility and awareness of the structure of the environment, are a more compelling case, as I've suggested above.

The biological world seems to me to be pervaded with powerful intelligence awaiting much deeper study than hitherto. In particular, we must not merely ask: what can such and such organisms do? We must ask what kinds of knowledge (or information) may be required for such performances? We must then ask in what kinds of ways such information may be represented? And we must ask what sorts of processes can operate on such representations to actually generate the observed behaviour with all its fine structure.

Boden says "purposive behaviour is not the same as behaviour controlled by feedback of the sort studied in classical cybernetics or control theory". While agreeing with the spirit of this remark, as indicated above, I think that such formulations risk inviting pointless boundary disputes. Instead of arguing with someone who replies "but they are essentially the same", we should co-operate in mapping out the terrain of the space of possible computational systems, so that we can be clear about the similarities and differences. In this case the differences imply for example that much of the mathematics developed in control engineering is irrelevant to designing or understanding the more intelligent purposive systems, for the reasons indicated above. Having got the differences clear, I don't think it is worth quarrelling over whether words like "purposive" apply to the simplest systems. (Map the terrain, instead of arguing about boundaries.)

UNINTELLIGENT HUMAN BEHAVIOUR

Besides intelligence in "lower" forms, there are aspects of "unintelligence" in "higher forms". Boden's discussion of insect behaviour refers to a lack of flexibility, including "patterns of invariant order which once started are automatically executed to the bitter end even in inappropriate circumstances". However, this phrase also characterises much of human life, especially at the level of social and international behaviour. But it also underlies many errors, from slips of the tongue to serious accidents. My own view is that there are deep reasons for this.

(a) Any computational system must at some level be based on mechanisms which merely behave, without deliberate planning, etc. Otherwise nothing could ever get started.

(b) The highest level goals, principles, evaluation strategies, etc. must be hard to alter (though alterable) if individuals are not to be erratic, unpredictable, whim driven and ill-suited to fitting into a social system.

(c) When the underlying mechanisms have speed limitations it will be crucial not to have always to derive plans and decisions from first principles, but instead to compile tried and tested solutions into rapidly executable subroutines and perhaps even to have them executed by lower level, intelligent subprocessors, whose behaviour may be hard to control or modify.

So, instead of clear boundaries between flexible, intelligent systems and relatively blind stupid systems, we must expect to find a judicious mixture of intelligence and stupidity, flexibility and rigidity in all systems.

Perhaps the central point is that the design of intelligent systems always involves compromises, trade-offs between space and time, generality and efficiency, efficiency and reliability, controllability and complexity of representation. To really understand animal behaviour, and evolutionary pressures, it is important to explore and understand these trade-offs. And that means doing A.I., i.e. studying the properties of computational systems. An example mentioned by Boden is the trade-off between modularity and simple control structure of production systems, and the need for more global control and richer representation of plans and strategies.

VISION

Visual perception remains one of the central challenges for A.I. and psychology. The power, speed, flexibility, and generality of function of visual systems defies explanation in terms of any current models, despite the advances mentioned by Boden. But at least we are beginning to understand some of the questions to be asked as a result of attempts at model building.

NAIVE PHYSICS

A quibble about the discussion of Naive Physics. Boden suggests that once natural language concepts are acquired, the meaning of the more primitive core concepts is altered. This presupposes a deep connection between the perceptual subsystems and motor control subsystems and the conceptual apparatus involved in language understanding and use. This (Whorfian?) hypothesis is by no means obvious, or even plausible. For many animals, and humans from many different cultures, seem to share a substratum of visual and motor abilities. A cat, a chimpanzee, and a human being are all able to walk through an opening, or drink liquid from a container for example. We don't know enough about the underlying representational abilities to be able to discuss what does or does not change as language is learnt.

METHODOLOGY

In "Problems of Experimental Validation" Boden mentions the very serious difficulty in studying a computational system from the outside, since exactly the same observed behaviour can be produced by infinitely many different programs (at least in principle). It is perhaps worth mentioning that it is hard to circumvent this problem by studying such systems from the inside, e.g. neurophysiology. This is because it is virtually impossible to "read off" high level programs from their physical representation, especially when there are many intermediate layers of "virtual machine".

There is another important source of methodological difficulty. An intelligent system composed of many interacting subsystems will not necessarily ever produce behaviour under the control of one subsystem. The actions of any subsystem are liable to be modified or even suppressed by the actions of others. Perceptual processes can be modified by motives and beliefs for instance. Language production and comprehension are clearly modified in a multitude of ways by internal and external context. It follows that even a correct model of some subsystem may often yield incorrect predictions of observable behaviour because the model cannot represent the functioning of the whole system. (The distinction between competence and performance, made by some linguists, is an attempt to grapple with this problem.)

In view of this difficulty, scientists must abandon the idea that their *primary* goal is to discover and explain laws with predictive power (whether deterministic or statistical). Instead the broader aim should be to understand the *generative potential* of systems, where a crucial feature of potential is that it need not be realised in predictable or controllable ways. (Compare chapter 2 of my book *The Computer Revolution in Philosophy* [1].)

MOTIVATION AND EMOTIONS

Boden's paper is mostly concerned with cognitive processes in the narrower sense, including perception, problem-solving, planning, and learning, but not "affective" processes like the formation of desires, conflicts of motives, moods, emotions, and the like. This, together with the fact that hardly any work in

A.I. has addressed such things, may lead some to assume that such processes cannot be understood in computational terms. Could a computer fall in love?

In order to investigate this topic in depth it is necessary to take our ordinary concepts "desire", "pleasure", "emotion", and a host of specific concepts related to these, and work out in great detail how we understand them, in order to clarify just what question it is that we are asking about them. Without such analysis people are often tempted after a moment's introspection to conclude that it is self-evident that they are not reducible to a computational analysis. Such temptations should be resisted. In particular, many ordinary emotion concepts refer to complex but analysable processes in which motives (of varying levels of generality) interact with other motives and with beliefs or doubts, to produce disturbances of cognitive processing. These are not always malfunctions, but may be essential for survival or successful execution of goals. For instance, the state of fear of falling off a narrow bridge which one has to cross is a state in which beliefs about what could happen, together with motives such as a desire not to fall, interact to produce a state in which the central processes are compelled to attend in detail to the problems of getting across safely. The system may of course over-react: there is a fine line between commendable extreme caution and debilitating fear.

Some emotional states involve very abstract and ill-defined processes and representations. For example besides specific fears there may be a belief that something nasty is likely to happen without any specification of what the nastiness is. Some emotional states, e.g. certain sorts of awe, need not even involve the belief that the awe-inspiring object is likely to do something nasty, merely that it has the capability of doing so. Thus it is possible to enjoy the security of being in the good books of a powerful agent whilst noting, and to some extent reacting to, the need to remain in her good books to avoid risking her wrath. Both positive and negative aspects of the relationship may interact with processes of perception, of decision making, of planning and execution. Only in relatively extreme cases would this sort of state be described as an emotion, but there is no clear dividing line, and we should again beware of fallacious boundary building.

The point of this example is that human emotional states may involve very complex mixtures of reactions to complex situations. The variety and complexity of such states, which in some sense we experience and recognise, yet which we cannot easily analyse, is one of the factors which make it appear implausible to suggest that computational systems could have emotions, and therefore that human beings and other animals are computational systems.

But this is based in part on a failure to survey possible computational architectures in depth. The existing sorts of programs, even the most complex ones, are far too simple to justify detailed comparison with human processes. For instance, their goal structures tend to form a simple hierarchy with a single goal provided by the programmer or user and all others generated as subgoals by a planning mechanism. An animal will, by contrast, have many "motive generators" concerned with a variety of needs of its own, its young, and the rest of its group. These needs will be of different kinds, and will not be static.

Moreover, in the case of human beings the process of absorbing a culture will normally include adopting a variety of high level motive-generators and motive-comparators, which can interact in complex ways. Instead of being awe-struck by this complexity and the difficulty of understanding what is going on, we need, as ethologists and cognitive scientists to analyse and explain the processes. And insofar as our analyses point to processes in which a variety of internal representations interact and produce new ones, which in turn cause further effects which may ripple through the system, they are computational processes. As such they can be replicated on a suitably powerful computing system. But first we need to find out what it is that is to be replicated, and that requires detailed task analysis.

CONCLUSION

I must apologise for not enlivening this journal with substantial disagreement! I hope that Boden's paper will inspire more researchers to adopt an interdisciplinary approach, using the information gleaned from empirical investigations both to inspire and to test deep theories about the computations underlying much animal behaviour. More trainee ethologists should be given the opportunity to study the concepts and techniques of A.I. There are many important but very difficult theoretical problems to be solved, and not enough good researchers are working on them. A.I. is in its infancy still.

REFERENCES

1. Sloman A. *The Computer Revolution in Philosophy: Philosophy, Science, and Models of Mind.* Harvester Press, Hassocks, Sussex (1978).