PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

A View of One's Own [and Discussion]

J. R. Lucas, M. Elton and A. Sloman

Phil. Trans. R. Soc. Lond. A 1994 **349**, 147-152
doi: 10.1098/rsta.1994.0119

A view of one's own

By J. R. Lucas
Merton College, Oxford OX1 4JD, U.K.

Two questions are distinguished: how to program a machine so that it behaves in a manner that would lead us to ascribe consciousness to it; and what is involved in saying that something is conscious. The distinction can be seen in cases where anaesthetics have failed to work on patients temporarily paralysed.

Homeostatic behaviour is often cited as a criterion for consciousness, but is not itself sufficient. As the present difficulties in surmounting the 'frame problem' show, ability to size up situations holistically is more important; so is the explanatory role of the concept.

Consciousness confers evidential status: if we ascribed consciousness to an artefact, we should be prepared to believe it, when it said its RAM was hurting, even though we could detect nothing wrong, contrary to our thinking of it as an artefact. A further difficulty arises from self-awareness and reflexivity.

Two questions arise when we consider the possibility of making conscious machines: one is the simulation problem, how to program a machine so that it behaves in a manner that would lead us to ascribe consciousness to it; the other is a conceptual problem, what is involved in saying that something is conscious. It is important to distinguish the two questions. Much discussion of artificial intelligence (AI) has been muddied by a failure to separate the technical considerations relevant to the simulation of conscious behaviour from the conceptual considerations of what is at stake.

Theories of meaning are largely to blame. Many philosophers in my time have suffered from theories of meaning which have rendered them unable to understand what other people are saying, or even to believe what they themselves know to be true. It would be quite out of place here to discuss in detail what has gone wrong in recent philosophical discussion: sufficient to warn non-philosophers among you to treat anyone who purveys a theory of meaning with the same suspicion as you would a used-car salesman. 'Hold on; don't believe a word he says, or you will be sold a pup.'

In the case of consciousness it has often been argued that the behavioural criteria are all that can be at issue, because, when we say something, we can only mean what the grounds are for our saying it. That this is false can be seen if we consider the gruesome case of patients who were administered an ineffective anaesthetic together with an all-too-effective dose of curare, which completely paralysed them, so that they were able to feel the surgeon's knife without being able to make it known to him that they were still conscious. They were conscious, though there were no signs of consciousness. Hence, consciousness is not constituted by the overt evidence of being conscious.

The conclusion has been resisted by some philosophers who argue that it is

Phil. Trans. R. Soc. Lond. A (1994) **349**, 147–152
Printed in Great Britain

TEX Paper

only by reason of the patients' subsequent testimony that we know of their terrible experiences. If the patients had died, or if they had been given a amnesiac drug that made them forget the horrible things they had undergone, then, so the argument runs, there would have been no fact of the matter, beyond the evident absence of overt indications of consciousness, and we should have no warrant for saying that they ever were conscious of the operative procedures being performed on them. But this is absurd, as well as callously inhumane. I should view entirely differently the prospect of an operation in which I was going to be genuinely unconscious from one in which I was going to suffer all the agonies of unanaesthetized surgery, even if later I was going to be made to forget it. More generally, verificationist arguments fail if over-extended. It is one thing to be sceptical about the meaningfulness of statements which are entirely disengaged from the rest of discourse, but it does not follow that we should be invincibly impenetrable to speculations about questions that, as it happens, we cannot decide. What happened before the Big Bang, did a lone dinosaur ever make its way to Oxford, are there other spacetimes entirely separate from ours? I do not know. In some cases I do not even know how we might come to find out; but I understand the question all the same. We need to distinguish the evidence on which we may properly make an assertion from what is involved in actually asserting it. The latter goes beyond the former. If I make a prediction, I stick my neck out. The grounds on which I make it are inevitably in the present or past, but what I actually assert is about the future, and if when the time comes, my prediction is not borne out by events, I am wrong, no matter how good my grounds were for making it. Equally with causal hypotheses. I always may be proved wrong by subsequent observations, even though my hypothesis had been very well supported by the evidence available at the time I put it forward. Similarly with consciousness, the criteria for ascribing consciousness are one thing, but what we mean when we ascribe consciousness is something more.

As Professor Slomson (this volume) points out, our concept of consciousness is not clear-cut, but is, rather, a cluster of concepts, with a corresponding multiplicity of of criteria: those used by an anaesthetist dealing with a human patient, those used by a biologist deciding to anaesthetize an organism before vivisection, those used by an ecofreak, asking a tree's permission before cutting it down for fuel, in addition to the two he mentioned, where we are concerned with whether someone is aware of something, and whether he is self-conscious. It is easy to confuse these different senses. Often the best remedy is to ask what the opposite is, what consciousness is being contrasted with.

In the sense of consciousness in which we wonder whether a robot may be conscious, three criteria are of prime importance: homeostasis, holistic assessment, and explanatory integration of behaviour. Homeostatic behaviour is a necessary, but not sufficient, condition of something's being conscious. We find it extremely implausible to ascribe consciousness to totally inert minerals: even the ecofreak does not ask the coal's permission before burning it. But the homeostatic behaviour of plants, like that of thermostats, is not enough to make us think they are conscious, and the doubts extend to the behaviour of animals moving away from noxious environments, towards food supplies, or engaging in reproductive activity. These may seem at first sight to be instances of conscious behaviour, but we withdraw the ascription of consciousness if subsequent investigation shows a machine-like insensitivity to the real needs of the actual situation. Dennett makes much of sphexishness, the apparently well-thought-out behaviour of a wasp making nests wherein to lay its eggs, which, however, seems to be an automatism, triggered by certain stimuli, even when wildly inappropriate (Dennett 1984). But Dennett's argument is two-edged. In seeking to rebut the ascription of consciousness to the sphex wasp, it allows that the apparently purposive behaviour constitutes some prima facie evidence in favour of the ascription, and in arguing that the ascription ought none the less be withheld in this case, it indicates another criterion for consciousness, which, if present, would strengthen the case for regarding an organism as, indeed, being conscious. In recent years workers in AI have been much exercised by the 'Frame Problem': they can program machines to behave in fairly complicated ways, but not to adjust themselves to the wide range of circumstances which they are likely to encounter. Organisms have evolved to be able – within reason – to adjust. Although they occupy some particular ecological niche, there is a much greater range of variation within that niche than any machine thus far constructed can accommodate itself to.

Thus far we have distinguished two marks of consciousness: homeostatic behaviour that suggests the agent has a goal which is pursued in spite of adventitious alteration of circumstance; and some plasticity of behaviour, revealing a sensitivity to a wide range of circumstance and an ability to size up the situation as a whole, and adjust behaviour to the whole of it, and not just a few salient features. A third mark of consciousness is its explanatory power. If we can integrate a wide variety of different pieces of behaviour as manifestations of a single state of mind, it is rational to posit that there is, in fact, a mind at work. Ryle gives a brilliant account of vanity:

> On hearing that a man is vain we expect him, in the first instance, to behave in certain ways, namely to talk a lot about himself, to cleave to the society of the eminent, to reject criticisms, to seek the footlights, and to disengage himself from conversations about the merits of others. We expect him also to engage in roseate daydreams about his own successes, to avoid recalling past failures and to plan his own advancement. To be vain is to tend to act in these and innumerable other kindred ways. (See Ryle (1949) and, earlier, Austen (1818).)

This makes sense if we think of there being some one who is inordinately interested in himself, but not otherwise. We can understand these diverse pieces of behaviour provided we posit the existence of a person, but not otherwise, and therefore it is reasonable to make this posit, just as it is reasonable to believe that electrons exist because by so doing we can explain many diverse phenomena. Leibniz gave *la liaison des phénomènes* as a reason for rejecting phenomenalism and believing in material objects as the reality that gave rise to phenomenal appearances, and the same line of argument justifies the ontological assumption that certain patterns of behaviour are to be understood as the behaviour patterns of conscious beings. As and when AI artefacts produce outputs that can best be construed as showing what they are up to, we shall have a further argument for thinking of them as conscious.

These three arguments are presumptive, not conclusive (Ross 1930; Hart 1948)†.

† The difference between the logic of necessary and sufficient conditions and that of defeasible argument and counter-argument is of fundamental, but largely unrecognized, importance in the philosophy of mind and in the humanities generally.

They can be defeated, as Dennett seeks to defeat the ascription of consciousness to the sphex wasp. Often it is the conceptual consequences of imputing consciousness that seem to tell most against making any such ascription, and it is useful to consider what follows from regarding an organism or machine as conscious. We are most keenly aware of the moral consequences: we think it wrong to cause animals pain, unless for a sufficiently good reason, whereas we have no qualms in cutting down a tree or boiling unanaesthetized carrots. If AI machines were conscious and sufficiently self-aware, we might feel it incumbent to consult them about their own future, perhaps even give them the vote. But these moral consequences are not fundamental: rather, they flow from a metaphysical view of conscious beings as centres, each having a view of its own upon the world. We should not cause people pain, because they are sentient beings, for whom pain is, from their point of view, bad. In attributing to them a point of view, we attribute to them a privileged authority on what that view is. I am the authority on me. What I say about me goes; not absolutely: I may be lying, dissimulating or malingering, or I may misunderstand the English language, or even misreport my own feelings; but these are only subsidiary derogations from my being generally the person peculiarly entitled to speak about me, and to have my words believed. I may have a terrible disease or a horrible wound, and the doctors may expect me to be in great pain: but if I assure them that I feel no pain, and my actions confirm that I am being truthful, then the doctors have to accept my word for it, and perhaps look for other explanations; perhaps I have had the operation under hypnosis, perhaps I have had accupuncture, perhaps I am one of those unfortunates who can feel no pain. Conversely, if I feel pain, I am to be believed, even though there is no discernible cause for it. It may be 'only psychological', it may be some malfunctioning of the nervous system, but it cannot be denied without imputing to me dishonesty. Similarly in the case of animals, except that there we have no distinction between behaviour generally and the special sub-class of linguistic behaviour: if we read their behaviour as that of a sentient being, we believe it in the absence, or even against the evidence, of physiological explanation; if a rabbit squeals, I think it is hurt or frightened even if I can see no cause of pain or ground for fear, whereas if a tyre squeals, I give it no probative force and disregard it as soon as I can account for it as the result of friction on the road's surface. If we came to regard AI artefacts as conscious, we should give more credence to their output than to their hardware, and be prepared to agree that something was wrong even though the most exhaustive examination failed to reveal it, provided the output, either by means of direct symbolism or through suitably modulated aversive behaviour, indicated that this was so. Once this point is reached, moral consequences follow: the artefact has a right to consideration, because it has a view of its own which cannot be subsumed under views available to us apart from its explicit avowals or significant actions.

It is part of the ideology of artefacts that they are entirely our creations and have no existence independently of what we choose to give them. But it is not clear that this condition is one that, in practice, must obtain. In an article some thirty years ago I specifically allowed that the day might come when we could create artefacts with minds of their own much as we can today procreate organisms with minds. But I argued then, and still argue today, that there were certain conditions on an entity being regarded as an autonomous being with a mind of its own. It could not be a Turing machine, or anything whose behaviour had been so programmed as to be entirely predictable by us. For then we should not need to posit it as a separate entity all of its own, but could see it simply as an artefact merely performing the manoeuvres we had instructed it to do. If I program my computer to emit a piercing bleep if it is moved, I do not think that it is in pain when I pick it up to reposition it. A reductive analysis eliminates the entity to which we can ascribe consciousness or a mind of its own. It is not just a case of 'origin chauvinism', as Dennett alleges, that makes us deny ontological status to entities that are Made, not Begotten, but a corollary of ontological parsimony. If we can give a complete explanation of the workings of a machine in mechanical terms, then there is no need to posit it as an entity on its own, and hence no justification for doing so. But if on the other hand we cannot account for the machine's output in terms of mechanical causes, but can understand it if we think of it as an autonomous entity, then we are justified in doing so, and in ascribing to it a view of its own.

The anti-reductionism required is a conceptual one, not a scientific one, though it carries with it some scientific implications. Very often in the sciences, as in other intellectual disciplines, we have occasion to introduce some concept which cannot be explained in terms of some simpler theory. I cannot explain entropy to someone who deals only with single systems; I have to have some notion of an ensemble or some concepts of probability. Although the chemist can investigate biochemical processes, the concept of an organism is one that lies outside chemistry itself: he needs a biologist to set the agenda, though once the agenda has been set, it is up to him to discover what the answers are†. In a more rarefied way, Tarski (1956) showed that the concept of truth could not be expressed within a formal mathematical system adequate for ordinary arithmetic. We can give partial explications, but not complete ones. In the case of AI simulations of conscious behaviour can be hoped for, with no definite upper limit to what may be achieved. If, further, the creators of new artefacts 'leave go', and produce machines that are not completely predictable, and can 'learn from their own mistakes', and 'consider their own projects, and how to improve them', then, provided their performance was convincingly good enough, they should not only have met our current criteria for being conscious, but we should have no conceptual resistance to regarding each one as having a mind of its own, capable of deciding for itself what it was going to do, and a view of its own, a representation of the world from its own point of view; in short that it was conscious.

The ability to adjust to unexpected variation of circumstance involves an ability to stand back and consider not only the circumstances but also the programs, what they can accomplish, and hence in what circumstances they are appropriate. Standing-backness and self-criticism are peculiarly characteristic of consciousness. Its first appearance among animals – the discovery on the part of tits how to peck through the tops of milk bottles, of rabbits how to gnaw through the plastic guards around growing trees, of household pets how to find their way home over long distances – attract notice and are commonly cited as evidence of animals being conscious like us. The full transition from being conscious to being self-conscious appears fairly late; only in human beings, and often not even then.

† I owe these points to H. C. Longuet-Higgins, originally in conversation in Cambridge 35 years ago, and more publicly in Kenny *et al.* (1972).

But the first origins of this transition seem to be deep in the concept of consciousness. The conscious organism is something separate from its environment, acting homeostatically so as to preserve itself, and reproduce its kind, capable, to some extent, of adjusting its strategies to fit the situation in which it finds itself, and hence with some power of reflection (Küppers 1990). And then we are in deep waters.

References

Austen, J. 1818 *Persuasion*, p. 1. London: John Murray.

Dennett, D. C. 1984 *Elbow room*, ch. 1, § 3. Oxford.

Hart, H. L. A. 1948 The ascription of rights and responsibilities. *Proc. Aristotelian Soc.*, 171–194.

Kenny, A. J. P., Lucas, J. R., Longuet-Higgins, H. C. & Waddington, C. H. 1972 *The nature of mind*, ch. 2, pp. 16–21. Edinburgh. (Reprinted in Longuet-Higgins, H. C. *Mental processes*, ch. 2, pp. 13–18 (1987).)

Küppers, B.-O. 1990 Information and the origin of life. tr. *Manu Scripta.* Cambridge, Mass.

Ross, W. D. 1930 *The Right and the Good*, pp. 19–20. Oxford.

Ryle, G. 1949 *The concept of mind*, p. 86. London.

Tarski, A. 1956 *Logic, semantics, metamathematics*, pp. 187–188, 247. OUP.

Discussion

M. Elton (*University of Sussex, U.K.*). Mr Lucas identified a gap between the evidence for consciousness and the claim that something is conscious. Does the behavioural evidence point to consciousness inside, or to future behaviours?

J. R. Lucas. The move from evidence to claim introduces a new concept, 'consciousness', which brings along other ideas such as 'agency' and 'responsibility'. The claim goes well beyond the evidence.

A. Sloman (*University of Birmingham, U.K.*). An engineering-design standpoint, too, can show this gap. There are many ways of producing the same behaviour: e.g. a look-up-table or a generative system. So many cognitive scientists and AI-researchers reject behaviourism. But suppose we know just how a system works, so that it is totally explicable. According to you, that system can't be conscious. Is this a matter of fact, or is it an ethical decision?

J. R. Lucas. What sorts of explanation are at issue? Some sorts have the effect of explaining away. If the design-level explanation you envisage explains away the phenomena of interest by using non-intentional language, then that system is not conscious. There could well be a form of explanation in AI which still required talk of goals, etc., so such an explanatory language would not be explaining away the phenomena.