Paper submitted to the *Journal of Consciousness Studies*
followed by the two mutually refuting reviews received and the author's reply to the Editor of *JCS*.

Scientific requirements for an engineered model of consciousness

David A. Booth
University of Birmingham, UK
D.A.Booth@Bham.ac.UK

Abstract

The building of a non-natural conscious system requires more than the design of physical or virtual machines with intuitively conceived abilities, philosophically elucidated architecture or hardware homologous to an animal's brain. Human society might one day treat a type of robot or computing system as an artificial person. Yet that would not answer scientific questions about the machine's consciousness or otherwise. Indeed, empirical tests for consciousness are impossible because no such entity is denoted within the theoretical structure of the science of mind, i.e. psychology. However, contemporary experimental psychology can identify if a specific mental process is conscious in particular circumstances, by theory-based interpretation of the overt performance of human beings. Thus, if we are to build a conscious machine, the artificial systems must be used as a test-bed for theory developed from the existing science that distinguishes conscious from non-conscious causation in natural systems. Only such a rich and realistic account of hypothetical processes accounting for observed input/output relationships can establish whether or not an engineered system is a model of consciousness. It follows that any research project on machine consciousness needs a programme of psychological experiments on the demonstration systems and that the programme should be designed to deliver a fully detailed scientific theory of the type of artificial mind being developed – a Psychology of that Machine.

Introduction

Computer scientists and engineers have long dreamt of building embodiments of consciousness based on silicon chips. Some important progress has been made, at least in clearing the ground. Current state of the art was illustrated in a special issue of this Journal in 2003 on "Machine Consciousness."

Discussion of such projects with neuroscientists has recently intensified in Europe. The European Commission's research programmes include the Factor-10 Project (Knoll & de Kamps, 2003) and one priority for their programme on Future and Emerging Technologies is the whole topic of next steps in machine intelligence, e.g. the previous "Beyond Robotics" and the proposed "Bio-inspired Intelligent Information Systems" listed at the http://fp6.cordis.lu.ist/fet/fetid-proplist.cfm page. Multidisciplinary workshops on engineered "models" of consciousness were held in Birmingham and Turin in the autumn of 2003, led by software and hardware engineers and scientists, and bringing in theoretical neurobiologists at Birmingham.

This paper points to limitations in those discussions as viewed by a multidisciplinary scientist and philosophy graduate, without training in computer hardware engineering or any skills in programming but with diverse experience in specifying the quantitative simulation of psychosocial and psychobiological sub-

systems (Booth, 1978a, 1988, 2002; Booth & Freeman, 1993; Booth *et al*., in press; www.qualityliving.org).

The strategies currently advocated within the community working on "Models of Consciousness" do not make use of principles that are well established in the philosophy of science. The community also makes no use of the scientific concept of a conscious process as implemented within cognitive experimental psychology. These two deficiencies are intimately related. This paper argues that failure to rectify both of them will have dire consequences for the enterprise of engineering conscious systems.

Building boxes *versus* building knowledge

The modelling of consciousness is currently dominated by an assumption that the procedure to follow is Design → Build → Test (DBT). This sequence is entirely appropriate to the practice of engineering: professional engineers must get on with producing dynamic structures that serve specified functions, using whatever science is to hand. However, such an approach is totally incorrect in engineering research: this is supposed to build scientific knowledge, not just to build physical or virtual machines. (DBT may not be peculiar to engineering research on consciousness but this Journal is not the place to widen such argument.)

Like all science, the engineering science of modelling consciousness requires empirical theory – in this case, theory that is capable of distinguishing between conscious and non-conscious processes. The scientific method is to build general theory by repeatedly cycling through the sequence Hypothesise → Test → Interpret (HTI). Existing theory is used to generate hypotheses that can be tested in adequately specified situations to yield interpretable results. A cycle is completed by going back to the theory, revised or not, to draw out further testable hypotheses. All stages of HTI cycles are fully operationalised and peer reviewed within the community of scientists having the relevant expertise.[1]

Scientific research into natural consciousness thus requires the building and testing of theory about the minds of human beings, or of members of other species on earth that may have some forms of consciousness. Therefore the process of improving the scientific understanding of artificial consciousness must include analogous research. A machine worth investigating should be used as what an engineer might call a "test bed" but for a scientific theory of the mind of that sort of machine, not for *ad hoc* assessment of the machine's functioning.

In other words, the engineer's sequence of Design → Build → Test needs to be nested within the Test stage of the scientific cycle, of Hypothesise → Test → Interpret and back again through theory development. The machine needs to be designed and built to provide (among other things) evidence for or against hypotheses inferred from adequately developed scientific theory of mental processes occurring in the type of machine under development.

Brains and bodies in environments

Engineering cannot be just the building of dynamic structures. The machines have to work, and to work well: that is, they must be adapted to their environment and

[1] This is not the place to review the philosophical details. The late Imre Lakatos is widely recognised as having used insights from Karl Popper, Thomas Kuhn and others to build a coherent philosophy of scientific method and the epistemology of empirical theory.

the tasks that it imposes on the engineered entity. Even for bridges, this includes the demands of human culture as well as the demands of the weather etc. The same is true of models of consciousness: the machine cannot be designed or tested without being so thoroughly networkable to its operating environment(s) that the social context is as much part of design and functioning as are the silicon chips and the communication and/or robotic interfacing hardware. This is necessary to the natural psychology of language, emotion, infant development, job-employee fit and *a fortiori* consciousness. It can be no less integral to a psychology of the machine.

That is, a key part of research in engineering science (and of any science of systems) is specifying the operational environment. The investigators of conscious processes in natural systems have to design and construct, or to select, situations in which the performance of tasks provides evidence that is important to the theory of how the system operates. They have the good fortune not to have to build and educate the human beings or the members of other species that they investigate.[2] Yet they still have to select their research participants to fit the design of situations and tasks to be observed, else the data are unlikely to engage adequately with the theoretically crucial hypotheses. Exactly the same, the investigator of artificial systems needs to design and to build an effective theory-developing unit, both the machine and also a range of hypothesis-testing scenarios and demands. Just building boxes to do tricks in DBT mode is not scientific research.

[2] Modellers of consciousness are lucky enough (or such a long way from their objectives as) not to face the ethical constraints yet that face breeding or even just using natural subjects for experiments. If engineers seriously believe that they are going to build conscious systems in the foreseeable future, then they need to start planning soon how they will deal with the relevant parts of the moral issues besetting research on young children, human embryos and laboratory animals.

That is to say, the modelling of consciousness will not be achieved by developing hardware and software that do interesting things, implement intuitive notions of different abilities or seem conscious to their creators (Aleksander, 1996). These enterprises can be great fun (and hard work) for the modellers and for those (like myself) interested in their efforts. Nevertheless, building fascinating machines does not in itself add to systematic knowledge.

The need for a scientific psychology of the machine cannot be avoided by limiting immediate ambitions to piecemeal modelling in the expectation that in due time we shall realise that consciousness has emerged from interactions among clever bits and pieces. This tactic fails to address the fundamental question for the whole enterprise: what counts as scientific evidence for the existence of consciousness, even in human beings, let alone in ants, bats, chimpanzees, dolphins, elephants, robots from a particular line of development or virtual machines with one sort of architecture?

Informal triggers to theory development

Anecdotes about amazing achievements by natural or artificial systems are used in discussions of models of consciousness to point to the subtlety and power of the processes that are sometimes involved. Anecdotes become genuinely productive, however, only after they have been used to drive the scientific cycle (theory-HTI-theory) and have resulted in improved understanding of the intelligence of the carbon or silicon system. If there is a theory (already extant, or to be developed from existing theory) that might account for the observed performance, then hypotheses need to be generated from it that are testable on systems and situations like that in the anecdote.

If a robot consisting of an arm and two cameras looks more realistic with ‘eyebrows’ over the lenses, what matters for science is not speculating on the parts of a human face that we should put on a robot to make it look good. The scientific

issues are, for example, what role our own conscious and/or unconscious mental processes play in that impression of the robot or attitude to it. For example, if this mentation turns out to be important in human self-identity or empathic experience, the science and the engineering can then interact to determine how a robot might be built to use its/her/his sight of human beings' eyebrows in the same way.

A more proactive approach is to develop scenarios that provide tasks for artificial systems to tackle. However, even a great number and variety of situations in which to develop the performance of machines would not help the modelling of consciousness unless each scenario were used to test hypotheses from current theory in a way that is publishable in the relevant scientific discipline.

In other words, until the modelling of consciousness adopts the relevant normal science (in Kuhn's sense), there is no basis for advancing scientific knowledge, let alone for provoking a paradigm shift like the forming of a scholarly consensus that a type of machine is in fact conscious. Einstein did not advance the physics of motion by ignoring Newton's precise and specific theory, nor Newton by ignoring Aristotle's.

The dependence of modelling consciousness on psychological science

In short, the engineer's dream of building a conscious machine will not be realised without constructing the most relevant new arm of engineering science. The engineering of motor vehicles differs from the engineering of bridges in requiring development of evidence-based theory of the performance of engines, road wheels, cabins and their connections in automotive machines. Analogously, unlike the engineering of communication and information technology, the engineering of conscious machines requires the development of evidence-based theory of the performance of perceiving, intending, reasoning and feeling systems. A great deal of such theory already exists for systems that we have long been able to investigate scientifically.

Consciousness is an aspect of the mind, at the very least in the case of the mind of a normal human adult. Therefore the research discipline that is key to models of consciousness is the science (*logos*) of mind (*psuche*) as recognised in the worldwide community of scholars, nowadays centred on the universities - the century-and-a-half-old academic area called Psychology.

The question whether or not a physical or virtual machine is in any way conscious involves a highly systematic set of empirical issues on which psychologists and behavioural zoologists are the experts. Effective tackling of such problems is outside the competence of academic disciplines not constituted by the study of individual minds in action. The scientific part-solutions existing for natural systems are already far more precisely realistic than the philosophically most sophisticated commonsense. As in all science and engineering, to solve the problems we need empirically dense theory to have been developed by much testing of key hypotheses against the sophisticated collection and interpretation of replicable observations, i.e. a basic science of consciousness of any sort of system in its home situation or ecological niche.

An education in the role of conscious and non-conscious processes in present-day scientific theory of human minds also inoculates against infection by still virulent pseudo-problems about ineffable private events that the later Wittgenstein debunked 60 years ago, such as treating phenomenological expressions as observational data (cf. Booth, 2004). Forty years before that, the Introspectionist school in American psychology was forced by its own "evidence" to recognise the fallacy of treating subjective experience as the primary source of scientific data on the workings of the

mind. That approach was refuted by experiences of unconscious processes "entering" introspectable consciousness, e.g. as flash memories. Although some cognitivists still succumb to the fallacy, the basic point keeps being rediscovered within normal scientific development. For example, details of social events "leave" consciousness in the amnesia for influences on one's actions (Nisbett & Wilson, 1977) that results from our thinking being directed by our theories of ourselves.

In addition, reliance on psychological science escapes the arbitrariness of politico-legal approaches, such as acknowledging the conscious (or personal) status of some robots when they become good enough members of human society or convincing enough prominent intuitivists (and maybe even the mysterians) by their professions of self-consciousness, i.e. attributing consciousness to others and counting themselves among them. (Animal behaviour experts reckon that pet dogs count themselves as part of the pack led by their owners but a robot seeking membership of human society would have to communicate this viewpoint linguistically and convince juridically and politically acceptable national or international commissions.)

It may well be correct ethically to regard such moves as a proper consequence of creating a conscious machine (if we ever did). Yet acknowledging this prospect fails to address the basic epistemological issue of how we could ever have fully rational grounds for believing a system to be conscious. Any number of anecdotes about the robot (or by the robot about us) would only flesh out an historical account of such a change in society. There would be no scientific understanding of what had happened and therefore no fundamental empirical rationale for accepting the new social conventions.

Psychology, neuroscience and social research

Scientific understanding of consciousness (as an aspect of the mind) not only is the task of psychology but also is that discipline's task alone. Neither neuroscientists nor social researchers can study conscious processes without using the relevant knowledge and skills of the current academic discipline of studying people's minds.

Brain imaging tells us nothing about the mental processes themselves that achieve conscious experience. It may tell us where some of the necessary synaptic fields are located but knowledge of what a metabolically or electrically active region has to do with consciousness depends entirely on the adequacy of purely psychological evidence on the performance of the possessors of the brains imaged.

Neuroscience, Evolutionary Biology, Economics, Political Science, Cultural Anthropology and other mechanistic and/or historical disciplines provide crucial explanations of the embodiment and/or the acculturation of the minds evidenced psychologically. They also add to the bases for speculating about the biological, social or ontogenetic origins of our species' capacity for mind. Yet these other disciplines are no more than background for what is now the professional province of those with a graduate level of education in psychological science and a research training in at least one of its major branches.

These specialisations of basic psychology include social psychology, biological psychology, developmental psychology and the study of adult physical perception, thought and action. All areas are cognitive in a broad sense but this latter area is called cognitive psychology in a narrower sense; this is rooted in the laboratory tradition within psychology, which was sustained in the face of American behaviorism in the second quarter of the 20th century by the basic and applied experimental psychologists in Cambridge UK (Broadbent, 1961, 1973).

Study both of the neural and somatic wetware and also of the historical culture on which each human mind depends tells us nothing of the reality that is the system of mental phenomena themselves or of how that type of causation operates. Similarly, inspection of the hardware or even the binary states of the compiled code in which the virtual machine is implemented tells us nothing of the algorithms by which the system performs successfully.

To explain the relationships between brain and mind or between culture and mind, or among all three, a quite different sort of theory will be needed from those of each of brain, culture or mind. The theory of these systems' or levels' relations cannot be causally mechanistic, in any manner like neurophysiology, psychology or economics. Probably such theory will have to be developmental, explaining the ontogenesis of each mind through successive gene-environment interactions (GxE) in the growing individual. This developmental psychology has to account for the increasing autonomy that arises from the earlier succession of GxE, at least when the individual has a healthy brain and normal social culture, even if with serious bodily or familial handicaps (like being blind or having been orphaned in infancy without effective fostering).

In consequence, the consciousness-engineering research programme itself has to be based on multidisciplinary developmental science, as some have recognised. The intelligent artificial system will have "innate" capacities but they will only be able to "grow" actual competence (e.g. in thinking and feeling) through education and training within social cultures and physical ecologies that are adapted to their capacities. Also these cultures will have to adapt as the capacities and competencies of engineered individuals and artificial-plus-natural functional groups increase and diversify. Furthermore, if there is relevance in human history and in communication and the use of tools in groups of other primates, once these developing artificial systems are making some progress within their niche(s) in human society, they may have to be left to interact with each other without further intensive interventions by engineers: freedom is a condition for intelligence, as well as the other way round.

It should be noted that neither this ontogenic development nor the cultural history of a new sort of system bears any mechanistic similarity to processes of evolution by survival of the fittest. Furthermore, despite (or because of!) the popularity of evolutionary explanations (and their connections also with the great hopes pinned on genomics), I doubt that phylogenetic biology, psychology or socio-biology will be of much use in the scientific hard graft of evidence-based construction of this 'inter-causal' theory. The main reason for my scepticism is the dearth of even observational data on evolution, let alone the infeasibility of controlled experiments, unlike the vast scope of investigations already started and envisagable that can feed into a multidisciplinary ontogenetic science of the life and minds of human beings and other terrestrial species and historical niches.

Psychology, philosophy and engineering

The mysterian approach to the "hard problem" of relating subjective experience to objective reality is to treat it as a genuine issue but beyond resolution by science and/or philosophy. Recognition of the relevance of psychological science takes away the main underpinnings of that position.

The brain/mind problem has been addressed by philosophical categories like supervenience (Chalmers, 1996), "non-reductive" reduction to classical or quantum physics (Ross & Spurrett, 2004), or the venerable conceptual apparatus for contrasting identity (with no coherent mapping within the supposed brain-mind) with dualism

(requiring the pseudo-causal notion of interaction between the mind and the brain or brain plus environs: Booth, 1978b). The empirical crudity of all these categories is shown up once psychology is acknowledged as an entire science autonomous of neuroscience (and of evolution, cultural history or introspection). Then the problem of consciousness can be recognised as amenable to solution but within science, not by philosophy. Thus, one necessary condition for the engineering of consciousness is sufficient advance in the scientific evidence on the workings of minds as such, especially human minds.

Larger ambitions in explanation and modelling of consciousness require good understanding also of the roles in the mind of the social and physical culture and of the brain and body. However, before it becomes possible to conceive of even these merely psycho-cultural or merely psycho-neural wings of the superordinate multidisciplinary science, each particular sort of organisation of reality must be recognised for what it is. As implied earlier, physiological, social and mental processes each have their own distinctive state-to-state causation, forming an internally complete network of causation of each sort of system, such as a human person, society or physiology. Scientific knowledge at each level has long been sufficient to preclude any mapping across such types of reality, whether from state at one level to state at another level (especially as a causal relationship) or from function to function (nomologically). This is a major reason to expect the superordinate theory to be non-causal, maybe ontogenetic.

The relevance of this vision now becomes more evident: operational understanding of the ontogeny across all levels of a natural consciousness is exactly the sort of science that engineers need in order to build a physical and virtual machine to operate on a solar planet within human society. In any case, it should be clear that the engineering of models of consciousness depends on some sort of scientific theory that encompasses the material and virtual machinery, the ecological and socio-economic functioning of the artificial systems and, equally crucially, their psychology.

Pre-scientific psychological constructs

The pre-history of scientific psychology is littered with empirically untestable labels for aspects of the human mind. St Paul had the flesh warring against the spirit. Mediaeval neo-Aristotelians had their four humours. Freud had an id, ego and superego.

In the late 19th century, the academic psychologists talked about “faculties” of the mind, such as cognition, conation and affect. These also were atheoretical constructs. They survive only as labels on journals, textbooks and undergraduate degree modules. The terms don’t do any actual science.

This is the weakness also in engineering that seeks to create “abilities” to think, attend, learn, and to have sensations, emotions, images, memories and so on. Perhaps the most advanced such use of constructs from psychological science is the “cognitive technology” of Haikonen (2003). Unlike the “cognitive technology” for analysing the minds of consumers (Freeman *et al.,* 1993), however, there is as yet no psychology which is experimentally testable on models using this sort of hardware. The terminology borrowed from psychology may help engineers to work out some differences between sorts of chip or of functioning machine. Be that as it may, the flaw in this sort of approach is that it does not help to build the scientific theory of the mental processes in a piece of cognitive technology that is needed in order to determine if any of the system’s achievements involve conscious processes.

It has been pointed out that virtual machines with any chance of being conscious will have to be constructed with complex "architecture," e.g. separate sets of reactive, deliberative and reflective processes (Sloman, 2002). I can't comment on whether or not this helps software engineers to distinguish different things that their programs should do. What I can see is that this approach has yet to move to making concrete advances in the theory needed to characterise specific and contentful algorithmic processes that account for the performance of the machine and hence to giving an opportunity for the structuring observations that could distinguish conscious from non-conscious mentation.

Psychological scientists themselves still sometimes use these unoperationalisable generic concepts as a substitute for genuine theory – for example, the "central executive" which is invoked to coordinate attention and effortful remembering. However, research papers using this term explicitly acknowledge that it is a cover for the current lack of testable theory on the management of concurrent processes of multiple types, of which there is considerable empirical understanding in each case but nothing much on how they interact (e.g., Baddeley, 1992). The standing scientific challenge is to break up that suspiciously autonomous and homunculus-like box into boxes having distinct functions but with some explicit processes for exchange or integration with other functions. Indeed, it has been argued (consistently with this paper) that there can be no central executive or any other generic function in the mind, because all cognitive processes are content-specific (Allport, 1980).

Finding meaning in the concept of a conscious process

This view of a science of consciousness and mental architecture fits with widely accepted philosophical conclusions about the relationships of terms in empirical statements to what is going on in the observable world (e.g., Quine, 1973). Very few (if any) terms in an empirical theory can be mapped onto objects in the world or even onto a delimited set of observational or experimental data.

This position on the science of minds, brains and societies also builds on the later Wittgenstein's debunking of introspection, without a tinge of the behaviourism that some read into his approach. Indeed, it was probably the behaviorist (*sic*) thinking at the time of which Wittgenstein (1953) wrote that psychology consists of experiments and conceptual confusion. He could have had in mind the then dominant behaviorists' concepts such as 'reinforcement' (Peters, 1954) and 'stimulus' (Hamlyn, 1957). These were failed attempts at physicalist reductions of intent (Anscombe, 1963) and percept (Hamlyn, 1957): they systematically traded on verbal ambiguity to hide the fallacy of sustaining a materialist monism in the face of any non-physical aspect of reality. Such philosophically crass ideas certainly did not afflict Bartlett's psychological experiments in the 1930s in the University of Cambridge (where Wittgenstein was also working). Nevertheless, the mentalist (now termed cognitivist) tradition that survived behaviorism was strongly infected by the introspectionists' concepts of the contents of conscious as a source of scientific data. Indeed, the introspectionist fallacy survives to this day among many non-psychologists when their research or practice touches on mental states or processes (cf. Booth, 2004).

It is quite wrong to suppose that Wittgenstein argued anything that should be called a denial of the existence of consciousness. His point was that, if we are aware of something, it is some aspect of the objective world that we are attending to, not the contents of some private world of awareness. Scientific study of the processes of attention may not be easy but Wittgenstein himself was very clear that there were legitimate empirical issues about the aspect of the world to which a person is

attending at any given moment: does the diagram look like a duck right now, or is it a rabbit? Indeed, the hermeneutic problem of determining if it *is* (meant to be) a rabbit may be insoluble in an art-form, as the deconstructionists claim. Nevertheless, what a zoologist's sketch means about the species can be constructed from the theoretical context, which itself can be tested against reality in a myriad of ways.

The Introspectionists' programme had collapsed into incoherence in the face of the phenomenology of memory, perception and action before the Behaviorists had begun to formulate their ideology. The field was left open to behavioristic denial of the objective existence of the mind because mechanistically interpretable analysis of observable performance was unconvincing. Only later did experimentally based theory of cognitive processes became sufficiently rich and precise to tie down the distinctions between different sorts of processes, including conscious and non-conscious processes.

As late as 1940, a famous report concluded that it was not possible to measure sensations, in the sense of the private experiences of stimulation by pressure, heat, tastants etc. At that time, the concept of sensation had yet to be adequately anchored into observationally based theoretical structures. Over the last 20-30 years, however, scientific theory of mental performance has had to develop, so that now it is feasible to generate testable distinctions between cognitive processes that do and do not involve conscious processes. Examples include mental rotation of three-dimensional figures, explicit recall of past material *versus* recognising the material as coming from the past but with no recollection of context, and being influenced by a sight while not being able to say anything about the sight. The psychologists' tests of hypotheses relying on the theoretical distinction between conscious and non-conscious versions of a particular process can be applied to distinguish between (a) perceiving an aspect of the situation (in the sense of acting successfully with regard to it) without any subjective experience of that aspect and (b) perceiving that aspect while conscious of it.

The relative recency of such capacity in psychological science illustrates why teams attempting to program abilities into machines need to include members having the education and training to test psychological theory on such systems.

Tests for specified conscious processes

It is a truism that processes within a complex system cannot in general be specified solely from relationships between outputs and inputs. However, science does not work by induction of explanations from observations.[1] Rather, hypotheses are deduced from the theory that can be used to design experiments or to select observations that test for processes which the data by themselves could not identify. If we have a rich enough theory of the processes in systems that are agreed to be conscious, we can look for evidence in input/output relationships of these or sufficiently similar systems for the distinction between conscious and non-conscious processing.

All that can be observed of a mind is its expression in relations between the system's outputs to the environment and its inputs from the environment. Certain relationships demonstrate an overt achievement by the system facing a task within a scenario. This is traditionally known in psychology as 'performance' or (most misleadingly) 'behaviour.'[3] Yet, in the cases relevant to consciousness, what is being tested is the presence or absence of content of subjective experience being necessary to explain the observed performance, as specified by a differential hypothesis embedded within a rich theory about how human minds actually work. This approach

can be extended to a rat's mind, an insect's landing gear and central nervous system, or an engineered model. In all cases, the prerequisite is that currently viable theoretical explanations include distinctions between conscious and non-conscious processes or some effectively equivalent distinctions among hypothesised information-processing mechanisms.

Footnote 3. Even for a radical environmentalist within psychology or the crypto-physicalist behaviourism of Hull and Pavlov, behaviour is not mere movement; it is control by stimulation. In plain English, action is always informed by perception.

Current scientific use of the concept of consciousness

As acknowledged above, for systems of interesting complexity, observed relationships between outputs and inputs are not sufficient to identify internal states. However, input-output functions are not nearly so limited as diagnostics if they are used within a rich and much-tested theory of the type of information-processing system under investigation. A fast-growing number of experimental paradigms yield just such results when the observations are interpreted within established psychological theory in terms of a distinction between awareness and unawareness. These can be illustrated by three examples from the cognitive psychology of the 1990s that operate on the borderline of conscious processing. These paradigms all depend on at least one stimulus and two responses, one specific to the stimulus and another that is evoked also by other stimuli (where 'stimulus' and 'response' simply mean patterns of input and output that are functional for the system). The results can substantiate the use of a distinction between unconscious and conscious processing, thereby giving a scientific use within a theoretical framework for a concept of being aware of or subjectively experiencing something.

Priming is now very widely used in many parts of experimental psychology. In this design, the first stimulus presented is so weak or brief that it does not reliably evoke its specific response. That stimulus is presented again, or another stimulus that evokes the same generic response as the first stimulus. That response to the second of the sequence of two stimuli will be faster (and/or more reliable) than the response to the same stimulus in a test in which the first stimulus was not presented.

The intriguing phenomenon of 'subliminal' perception has now been tied down by measurement of the detectability (d') parameter of signal detection theory of the first stimulus by its specific response and by the generic response (Merikle & Cheesman, 1987; Merikle *et al*., 2001). There is subception of the first stimulus when it is not detected by its specific response (i.e. was unconscious, if perceived at all) but is detected by the generic response, i.e. it was indeed perceived, but unconsciously.

Thirdly, the discrimination of differences in strength of the first stimulus by gradations in its specific response can itself be differentially discriminated (i.e. influenced quantitatively) by gradations of a generic response. When this relationship between a stimulus and its response is better discriminable by the generic response than is either the stimulus itself or the conceptual state controlling the specific response, then there is "deeper" discriminative processing of information. This might correspond to a subjective experience such as a sensation or an emotion, if one also makes the testable hypotheses that (for sensation) best discrimination of the stimulus, when it occurs, is unconscious or (for emotion) the generation of the specific response is without experienced affect (Booth & Freeman, 1993).

These distinctions within priming, detection and discrimination are drawn for all sorts of cognitive process, not just the perception of physical situations. For any cognitive process to be diagnosable, its input has to be influenced by external stimulation. The stimuli can be verbal, or symbolic in some other way, e.g. the

statement of a premise in reasoning or the past experience of a choice between an action and an alternative (remembered intention). The process must also influence some output, but this response also can be symbolic rather than concrete: that is, the output can be statements that differentiate between situations; it does not have to be physical acts. The paradigm known as 'priming' has been applied to memories (perception of past situations), images (perception of situations not currently sensed), emotions (perception of social situations) and thoughts (perception of potential dialogue).

In work on implicit and explicit memory, for example explicit recollections or implicit memories as in expectations (Stacy, 1997) or reasons for actions, intentions, decision making or the Will (Frith, 1990), the usual methods seek qualitative differences or measure detection. Discrimination between difference levels of input can also be used. Such an experimental analysis would end the looseness of the protracted discussions of Libet's sign of pre-conscious decision and would show whether or not the evidence on the sequence of conation meets the scientific criteria that the discussants have assumed inituitively.

The same applies to artificial systems that are claimed to have abilities such as having mental images, experienced emotions, contentful thoughts and meaningful uses of a human language (Haikonen, 2003). The thinking or the emotion has to be affected by some external stimulus, but this can be verbal or symbolic in some other way and so the paradigms are not limited to perceptual processes.

Even self-consciousness is amenable to scientific investigation once the theory of cognition is rich enough. Once cognitive psychological theory has been developed under the challenges set by well designed sets of experiments or observations,[4] it will be feasible to determine the environmental contingencies under which a neural sign of self-consciousness is seen (Taylor, 2000).

Footnote 4. Note that of course this requires research programmes, not anecdotal observation or one-shot experiments.

Adaptation of natural psychology to artificial minds

The processes invoked by the scientific theory of natural minds are unlikely, however, to transfer to the design of artificial minds without considerable adaptation. The virtual machine in the human person is implemented in algorithms (i.e., cognitive processing), code (e.g., highly distributed patterns of axonal and dendritic potentials and thoroughly inexplicit non-verbal communication and informal cultural dynamics) and processors (i.e., sparsely adapting but richly interconnected dendritic trees and interlocking institutions with indeterminate openings for creativity and criminality) that will never be mimicked in silicon or molecular (nano) technologies or internet transfers.

Indeed, it is a misconception of the nature of mind to attempt to engineer machines that are conscious by simulating some aspect of the human being. Like any other machinery, computing systems should be designed to do what is needed as well as possible with the hardware and software available. When identifiable artificial systems begin to play intelligent roles in human society, then scientific investigation of machine consciousness can begin. The relevant parts of then-current psychological theory can be adapted to the physical embodiment and the social context of a system and hypotheses derived that could provide interpretable data on conscious processes from the system's performance in appropriately designed scenarios. The results will instigate further adaptation and elaboration of the psychological theory to this 'species' of model. We may then begin to understand the sense in which those

systems are conscious and also various ways in which their consciousness is similar to ours and different from it.

A research programme on engineered models of consciousness is unlikely to be able to use the theory of the minds of human beings, or of rats, bees, chimpanzees and other animals, exactly as it exists. By definition the hardware is very different. AI research and computational psychology soon learnt that the programming language and operating system of a natural mind is very different – so different that it is not clear if these computing entities have any direct homologue in animals. Less obviously perhaps, the public life of an engineered intelligence is liable to be very different from that of a natural mind. To deploy a biological concept, the ecological niches of robot and distributed systems will be quite distinct from those of human beings. In sociological terms, the role of an engineered member of society would be authentic to itself, not authentically human.

Whatever empathy we may feel for the engineered minds (or believe they feel for us) will be practically important. Yet our knowledge of their consciousness will be based solely on objective evidence on how they perceive their environment, their actions, the human beings with whom they interact and themselves (Broadbent, 1973).

Research programmes

It should be clear from all of this that these challenges from natural to artificial modelling of consciousness are multi-layered. Minds and ecologies had to co-evolve in nature. This makes it very likely that engineering scientists can develop conscious machines only by developing the social and physical environments in which the engineered entities operate and by understanding at an algorithmic level how the machines cope with the tasks faced.

One implication is that the conscious machine will have to perform genuine tasks within human society. This is not just a matter of *realpolitik* and economics. It is an epistemological and scientific necessity. There is no way of rationally attributing consciousness to the machine except by establishing a biosocial cognitive-behavioural science of its life and times by the cycle of theory construction, hypothesis, test and interpretation into further theory.

A corollary is that, even for the purpose of advancing theoretical knowledge, straight engineering efforts would best be put into the building of systems that serve practical functions that previously only human beings could perform. The usefulness of these machines would not merely be a help in getting the research funded. Achieving worthwhile functionality makes it more likely that systematic investigation of the system's performance can be relevant to a science of artificial consciousness. That is even more likely if the systems also, for example, manage unusual challenges, repair breakdowns and automatically develop more sophisticated functions, as creative human beings do by bringing their whole minds to bear, conscious and unconscious.

This is quite different from an argument that the research programme will have achieved something worthwhile even if consciousness does not emerge. The present argument is that, whatever the outcome, scientific experts on natural consciousness will have to have been involved: the only way of testing for consciousness is to adapt existing theory to the system's range of achievements and then to construct scenarios and tasks within them that provide evidence as to whether the system is aware or not of some particular aspect of the situation while the task is being tackled.

The immediate practical implications of the argument are quite mundane. It will not be sufficient for computer scientists and engineers merely to educate themselves

in psychological science, even to graduate level. It is essential that scientists with professional research training and continuing productivity in the purely psychological understanding of natural systems are integrated into the engineering of consciousness.

This integration could include multidisciplinary individuals who publish research into the cognitive performance of both robots and human beings. More likely, existing hardware and software engineering scientists will work with contemporary social, biological and cognitive psychologists and other behavioural scientists in joint research programmes. Such teams could build both increasingly sophisticated artificial systems and also new theory that distinguishes conscious from non-conscious processes in the interactions of those systems with the environment.

None of this will start until at least a few established workers start running publishable experiments on the psychology of the machine and attracting young entrants to this so far unrecognised research field.

References

Aleksander, I. (1996). *Impossible minds: my neurons, my consciousness*. London: Imperial College Press.

Allport, A. (1980). Patterns and actions: cognitive mechanisms are content-specific. In Claxton, G. (Ed.), *Cognitive psychology: new directions*, pp. 26-64. London: Routledge & Kegan Paul.

Anscombe, G.E.M. (1963). *Intention*. 2nd Edition. Oxford: Blackwell.

Baddeley, A. D. (1992). Is working memory working? *Quarterly Journal of Experimental Psychology* 44A, 1-31.

Booth, D.A. (Ed.) (1978a). *Hunger models: computable theory of feeding control*. London: Academic Press.

Booth, D.A. (1978b). Mind-brain puzzle versus mind-physical world identity. [Comment on R. Puccetti & R.W. Dykes, Sensory cortex and the mind-brain problem] *Behavioral and Brain Sciences* 3, 348-349.

Booth, D.A. (1988). A simulation model of psychobiosocial theory of human food-intake controls. *International Journal of Vitamin and Nutrition Research* 58, 55-69.

Booth, D.A. (2002). *evidence-networking Application of Best Life Education research (enABLEr)*. http://eoi.cordis.lu/dsp_details.cfm?ID=28321 or via link on www.qualityliving.org - *since 2007*, www.wwiyc.org

Booth, D.A. (2004). Phenomenology is art, not psychological or neural science. [Comment on S. Lehar, Gestalt isomorphism and the primacy of subjective conscious experience: a Gestalt Bubble model] *Behavioral and Brain Sciences* in press. (*Since published at* 26, 408-409. DAB's riposte to Lehar's reply to Comment published on Lehar's webpages for this BBS commentary at http://cns-alumni.bu.edu/~slehar/webstuff/bubw3/BoothResponse.html)

Booth, D.A., & Freeman, R.P.J. (1993). Discriminative measurement of feature integration in object recognition. *Acta Psychologica* 84, 1-16.

Booth, D.A., Blair, A.J., Lewis, V.J., & Baek, S.L. (2004). Patterns of eating and movement that best maintain reduction in overweight. *Appetite*, under revision

Broadbent, D.E. (1961). *Behaviour*. London: Methuen.

Broadbent, D.E. (1973). *In defence of empirical psychology*. London: Methuen.

Chalmers, D.J. (1996). *The conscious mind: in search of a fundamental theory*. New York: Oxford University Press.

Freeman, R.P.J., Richardson, N.J., Kendal-Reed, M.S., & Booth, D.A. (1993). Bases of a cognitive technology for food quality. *British Food Journal* 95 (9), 37-44.

Haikonen, P.O. (2003). *The cognitive approach to conscious machines*. Exeter: Imprint Academic.
Hamlyn, D.W. (1957). *The psychology of perception: a philosophical examination of Gestalt theory and other theories of perception.* London: Routledge & Kegan Paul.
Knoll, A., & de Kamps, M. (2003). *Roadmap of Neuro-IT Development*: http://www.neuro-IT.net
Merikle, P.M., & Cheesman, J. (1987). Current status of research on subliminal perception. *Advances in Consumer Research* 14, 298-302.
Merikle, P.M., Smilek, D., & Eastwood, J.D. (2001). Perception without awareness: perspectives from cognitive psychology. *Cognition* 79, 115-134.
Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we know. *Psychological Review* 84,
Peters, R.S. (1954). *The concept of motivation*. London: Routledge & Kegan Paul.
Quine, W. V. (1973). *The roots of reference*. La Salle: Open Court.
Ross, D., & Spurrett, D. (2004). What to say to a sceptical metaphysician: a defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences* in press.
Sloman, A. (2002). Architecture-based concepts of mind. In G. Hatona, N. Okada & H. Tanabe (editors), *Affective minds*, pp. 169-181. Elsevier, Amsterdam.
Stacy, A.W. (1997). Memory activation and expectancy as prospective predictors of alcohol and marijuana use. *Journal of Abnormal Psychology* 106, 61-73.
Taylor, J.G. (2000). Attentional movement: the control basis for consciousness. *Society for Neuroscience Abstracts* 26, 231 (839.3).
Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell, Oxford.

[E-mail correspondence about the above MS follows.]

REVIEWS accompanying the Editor's decision to reject

David A Booth:
Scientific requirements for an engineered model of consciousness

Review One

The author proposes that in order to assess whether intelligent machines have consciousness it will be necessary to involve psychologists who have experience using theory-based approaches to understand consciousness in humans. There are a number of assumptions underlying this position. Two of the most important are a) that "contemporary experimental psychology can identify if a specific mental process is conscious in particular circumstances" (p. 1) and b) that "observed input/output relationships can establish whether or not an engineered system is a model of consciousness" (p. 1). In my opinion, both of these assumptions are questionable because they imply that it is possible to use third-person data both to identify conscious mental processes and to distinguish conscious from unconscious processes. I don't know of any experimental psychological approach to the study consciousness based on third-person data (e.g., discrimination, detection, RT) where the behavioural measures have not been first validated by subjective experience (i.e., first-person data).

Perhaps my criticism is incorrect. One way to address the criticism is to discuss this issue more thoroughly in the paper. In other words, are there any situations in which measures of consciousness based on third-person data have been validated independent of any reference to subjective experience? I can not think any. If there are any such situations or behavioural measures, I would certainly like to know about them. Another way to approach this issue would be to discuss Searle's Chinese Room thought experiment. The position put forward in the paper is clearly at odds with Searle's arguments against a strong Al position.

Another issue that is important to address concerns whether or not consciousness is causative. There is the hope expressed in the paper that it will be possible to distinguish "conscious from nonconscious causation in natural systems" (p. 1). The basis for this optimism is unclear to me. I don't know of any research findings in experimental psychology which demand an explanation in terms of consciousness. Rather, all the findings to date could easily be simulated by machines without any reference to consciousness. Thus, there are few if any empirical reasons for rejecting the idea that consciousness is an epiphenomenon.

Given these criticisms, it's not at all clear to me why psychologists would be particularly successful at determining whether intelligent machines have consciousness. The only reliable psychological data regarding consciousness are first-person data based on reports of subjective experiences. Given that first-person data are always used to validate third-person data, I don't see how psychological approaches will necessarily provide insights into how to determine whether intelligent machines have consciousness.

Review Two

About the paper in general: This paper is an opinion paper. This is not a scientific paper; it does not introduce any research results nor any new ideas or hypotheses for future research. This is not a philosophical paper, either; no philosophical ideas are developed or elaborated here. The title of this paper "Scientific requirements for an engineered model of consciousness" is misleading as no scientific requirements are actually presented here. The only actual message of this paper is that psychologists (i.e. David Booth himself) should be involved in research programs on machine consciousness.

The author tries to convince the reader by the logically invalid argument by authority. In the introduction chapter the author positions himself as a multidisciplinary scientist, however without any training in computer hardware engineering or programming. Many of the references are only given for the purpose of proving the competence; the content of these references are not related to (engineered) models of consciousness.

The chapter "Building boxes vs. building knowledge" is constructed around naïve and misunderstood view of engineering process. The author proposes that design engineering were based on the procedure design-build-test. This is not the case. Engineers cannot design anything without a theory. The actual procedure is iterative and more like: develop theory - design experiments - build, test - verify and/or modify theory - design and simulate actual equipment - build - test - and perhaps start again. This procedure is also used in engineering efforts towards machine consciousness.

In the chapter "Psychology, neuroscience and social research" the author states: "Similarly, inspection of the hardware or even the binary states of the compiled code in which the virtual machine is implemented tells us nothing of the algorithms by which the system performs successfully." This is not true. The correct or incorrect functioning of a program, even the programmed algorithm itself, can be determined by the inspection of binary states of the hardware; there is even a specific instrument for that, namely the logic analyzer. This kind of work may often be tedious, though, but that is beside the point. With this fallacious argument the author tries here to argue that the hardware or neural wetware were separate from the mental phenomena.

The paper contains an overall problem; the concepts of mind, consciousness, self-consciousness, self-history and the contents of consciousness are frequently mixed. This is manifested e.g. in the sentence: "There is no way of rationally attributing consciousness to the machine except by establishing biosocial cognitive-behavioural science of its life and times..."

The author argues against the conventional concepts in psychology like cognition, conation and affect and maintains that these "survive only as labels on journals, textbooks and undergraduate degree modules. The terms don't do any actual science". It is true that labels only do not do science; e.g. we have not explained anything if we say that a gadget amplifies because it is an amplifier. However, I suspect that modern psychology does have something in the way of empirically grounded structure behind the labels and I think that this view is shared by many. Why does the author deny this? The answer is not evident from the paper, but can be found at Prof. Booth's web site. Prof Booth subscribes to "Individualised cognitive analysis".

According to Prof. Booth's web site "Individualised cognitive analysis provides direct evidence as to what is going on in a person's mind (or in any well adapted system's performance) while tackling a task such as recognising and acting appropriately towards an object, a social situation, an emotional state or a bodily sensation. The evidence can be purely verbal, from an adequately structured conversation, or can be concrete actions or expressed dispositions in response to physically defined stimuli or culturally meaningful symbols (such as words or pictures). My approach is to compare the person's responses to variants of the situation under test that disconfound features from each other and from their context. The data from one test occasion are analysed by multi-channel discrimination scaling: this is the simplest formulation of the classic ideas of dimensions of mental processing, learnt Gestalten and the just noticeable difference, and in that sense forms the logical foundation of all psychology." (sic).

In this way Prof. Booth effectively denies much of the mainstream psychology.

The paper as a whole does dot clearly indicate what superior tools and methods psychology could offer to the engineering of machine consciousness; the author already admits in the abstract that "Indeed, empirical (psychological) tests for consciousness are impossible because no such entity is denoted in the theoretical structure of the science of the mind, i.e. psychology." This does not prevent the author from demanding in the final chapter: "It is essential that scientists with professional research training and continuing productivity in the purely psychological understanding of natural systems are integrated into the engineering of consciousness."

However, would David Booth's "Individualised cognitive analysis" be the answer as the author obviously is implicating? More exact arguments, substantiated by empirical evidence, would be needed here. First the author should convince the reader that the "Individualised cognitive analysis" is indeed a valid method and does replace conventional approaches and then the author should prove that his method really tells something about consciousness and provides something beyond anecdotal verbalised introspection. This is not done here. This is one reason why this paper falls into the category "opinion".

Author's reply to the Editor

Dear Editor

(A) These two reviews flatly contradict each other.

One: "The only reliable psychological data regarding consciousness are first-person data based on reports of subjective experiences."

Two: " ,,, the author should prove that his method really ... provides something beyond anecdotal verbalised introspection"

You can draw any conclusion whatsoever from a contradiction - rejection or acceptance - or revision - or other reviewers.

(B) Both reviewers attack the paper by mere fiat ("opinion").

One: "discrimination [or] detection" (e.g. of sugar in a cup of tea) by a first-person expression of subjective experience (like "It seems sweet to me") is classified a priori as "third-person data".

Two: my method of "discrimination scaling" (e.g. of sugar in a cup of tea) by "anecdotal verbalised introspection" (like "It seems sweet to me") does not count as "empirically grounded structure" (despite the publications in psychological and other scientific journals cited on my website).

(C) May I suggest therefore that it could be very unfortunate to reject a paper on the basis of an incoherent display of presuppositions about the status of the concept of a conscious (or unconscious) cognitive/mental process within psychological science of the last 20 years? It's hard to believe that either of these reviewers is acquainted with the current discipline, even as little as the engineers and philosophers (of consciousness) who have commented on the paper.

Best wishes to the Journal.

Author

On 27 Jul 2004 at 15:42, Anthony Freeman wrote:

Date sent: Tue, 27 Jul 2004 15:42:58 +0100
To: David Booth <D.A.Booth@bham.ac.uk>
From: Editor <Editor@imprint.co.uk>
Subject: JCS Submission

Dear Mr Booth

I have now received two substantial reviews of your paper on "Scientific requirements for an engineered model of consciousness". Neither referee advises acceptance of the paper for publication in this journal and the editors are following that recommendation.

For your information I attach the two reviews, but this should not be construed as an invitation to resubmit a revised version of the paper.

Yours sincerely,

Editor