# Towards an Axiomatic Theory of Consciousness

Jim Cunningham, Imperial College, London SW7 2BZ

rjc@doc.ic.ac.uk, March 2000

## Pragmatics of Consciousness

Although part of the ancient mind-body problem of philosophy, the concept of consciousness itself is well enough recognised for it to be an ordinary word of our language. A conscious individual is aware, and knowing; the unconscious condition is normally recognisable. Yet numerous popular and contemporary books by Searle, Dennett, and others, show its explication to be contentious and a challenge to our suppositions on reality; a hazardous topic indeed for a would-be engineer of artificial intelligence.

Our justification for addressing the subject is that artificial agents which display elements of intelligent behaviour already exist, in the popular sense of these words, but that we would doubt the intelligence of an agent which seemed to us to have no sense of "self", or awareness of its capabilities and its senses and their current state. So an approximation to human consciousness could enable us to converse more naturally with an individual agent. We do need an account for the first person perspective as well as the second and third, and we cannot rule out the possibility that consciousness has utilitarian function, evolved to ensure survival.

Contention arises over whether consciousness can be considered a mental state of the human mind, for this brings presuppositions of the intentional stance and issues of faithfulness to the human model. But lack of faithfulness to a biological model is not a barrier to engineering, as the wheel, the fixed wing, and the computer itself demonstrate. Software agents are already designed with mental states and practical reasoning methods which have emerged as abstractions from rational enquiry rather than any physical brain model. While agent designers may also eschew such models, and instead rely on a variety of physical and computational devices, in well known cases the management of complexity leads to design architectures with layers of abstraction, some of which are comparable with intentional models of the mind.

To bypass the metaphysics of consciousness in favour of pragmatic considerations, there is evidence to consider, the view of peers in rational enquiry, and the need for guidance in an artificial construction. Clinical reports, psychological experiment, and philosophical enquiry, lead to a variety of theories which partially explain the phenomena and suggest layers of consciousness. Problematic issues range from neurological phenomen such as phantom limbs and the relation to wakefulness and unconscious mental processing, through issues of identity and the effect of emotion and the habitual, to an explication of context and presence in perception and of the links with language and intentionality.

Our formalism arises from attempts to bridge the gap between agents designed with mental states, and credible multi-processing implementations. It may be compatible with the implementations of a psychologically motivated theory like that of Baars, which can be realised as a computational agent with a myriad of heterogenious processes. But to explicate conscious behaviour we may still require layers of conception which we hardly discuss here.

## A Refined Intentional Stance

Mental models of the intentional stance encroach on two areas of agent design. One is as an abstract basis for incorporating plans and the selection of actions through means-end reasoning in software agents, notably in variants of Rao and Georgeff's Belief, Desire, Intention (BDI) paradigm. The other area is the related basis for giving definition to standard acts of communication as realisations of speech act theory, so that there are ingredients of a coherent basis for dialogue between agents in terms of what we can loosely call knowledge interchange.

Our proposal for steps towards an axiomatisation of consciousness depends critically on a refinement of traditional ideas of intentionality. From the perspective of an agent designer, extant intentional theories of rational agents focus on *stative* concepts like the BDI concepts themselves, and *knowledge* and *commitment*, each of which can be regarded as expressing computational *data* states. Activity, or process states, which are equally important in a computational model, have been ignored, or rather buried in naive computational models. But activity states like *planning*, *learning* and *sleeping*, and the *sensing* and *perceiving* of external conditions are equally important for a computational model of a rationality. This is a serious deficiency in the usual perception of

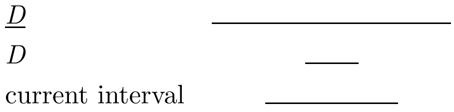


Figure 1: interval relations $D$ and $\underline{D}$

mental state, and we suggest a remedy below. However, there is another defect. The usual axiomisation for belief, and of knowledge, presumes introspection; e.g. for knowledge, that which is known is known, that which is not known is known to be not known. These are strong conditions which make such states already too "conscious" for some forms of memory recall and learnt behaviour.

The limitations of stative mental states can be overcome by simply allowing activity states as well. Each class can be considered durative on a temporal frame, but we gloss over the homogeneous / compositional distinction by using the progressive form of the activity. We capture this by a modal operator *prog* to modify the singular verb predicate, so that, for example, a rendering of *j is sensing c* becomes $prog\ senses_j\ c$. To define the *prog* operator we use the interval temporal logic of Halpern and Shoham, which uses modal operators to incorporate the interval relations identified by Allen. The during relation **D**, and its complement $\underline{\mathbf{D}}$ with respect to the current interval are illustrated in figure 1. We may define *prog p* to be the coercing form $\langle \underline{D} \rangle [D] p$, read (right to left) as *p holds on all sub-intervals during some interval which contains the current interval*. The usefulness of the Halpern and Shoham logic for representing and modelling both aspect and tense is demonstrated in a thesis by Leith and is being explored in joint work with Lloyd Kamara.

Once we have the ability to express temporal relations between interval based activities as logical properties, the interactions between activities and other mental states can be expressed by axioms. We may for instance consider that an axiom like:

$$\begin{aligned} &prog\ perceives_j\ p \leftrightarrow \\ &\qquad prog\ senses_j\ c\ \wedge\ prog\ remembers_j\ (c \rightarrow p) \end{aligned}$$

expresses the idea that sensory perception amounts to an ongoing inferential interaction between autonomous sense and memory recall processes. Although we could be more precise by constructing a compound sense process from realisations of particular sense mechanisms, the more problematic element of this definition is memory recall itself. Here we postulate a subconscious process rather than a stative belief state, because as mentioned above this would already be too strong a condition. It seems too that we have capability for richer expressions of tense and aspect which could encapsulate aspects of learning.

## Introspective Awareness

The perception processes posited above could be those of a sophisticated but unconscious automaton. The awareness needed for consciousness can also be considered a mental process rather than a data state, one which includes sensory perception, but also meta perception of sensory perception. Awareness can also be switched on and off by paying attention, either in response to change in sensory perception, or through volition; primitive processes whereby mental activity and ultimately action are controlled. We follow Carl Ginet by arguing not only for such philosophical abstractions but because mental control processes relate to notions of will and causality through neurological elements such as the motor cortex.

Thus to introduce consciousness, and ultimately a consciousness of responsibility, we need the activity of being aware to be positively introspective and controllable to some degree. When an agent is aware, it not only perceives, but at least for some conditions, perceives that it perceives. Thus assuming positive meta perception only, we might provide an axiom for a progressive form of introspective awareness as:

$$\begin{aligned} &prog\ aware_j\ p \leftrightarrow prog\ perceives_j\ p \\ &\qquad\qquad \wedge\ prog\ perceives_j\ prog\ perceives_j\ p \end{aligned}$$

This may be unnecessarily clumsy, not strong enough, and obscuring distinct perception processes, but because the scope and degree of introspection can be graded there seems to be no evolutionary argument against the acquisition of such higher levels of perception, indeed it seems necessary for a sense of social responsibility. A socially conscious agent which perceives a causal relationship will also perceive consequences of its perception of this relationship. We claim that once an agent has mental activities of sufficiently introspective awareness it also has a form of consciousness, that in its weakest form consciousness is just a progressive activity of being introspectively aware of something: $\exists p.prog\ aware_j\ p$. Graded and focused consciousness can follow.

## Bibliography


B.J.Baars,*In the theater of consciousness*, OUP1997

M.Bratman,*Intention, Claims and Practical Reason*, Harvard University Press 1997

D.Dennett, *Consciousness Explained*, Penguin 1991

C.Ginet, On Action, CUPress 1990

M.F.Leith, "Modelling Linguistic Events", PhD Thesis, Imperial College, University of London 1997

A.S.Rao and M.P.Georgeff,"BDI agents:from theory to practice", ICMAS-95, San Francisco, CA, 1995

J.Searle, *Mind, Language and Society*, 1999 (Orion)