Towards an Axiomatic Theory of Consciousness

JIM CUNNINGHAM, *Department of Computing, Imperial College, 180 Queen's Gate, London SW7 2BZ, UK. E-mail: rjc@doc.ic.ac.uk.*

Abstract

In this paper we seek to provide elements of an axiomatic theory for a sentient consciousness as a quantified form of introspective awareness. A crucial step for its formulation is use of an interval temporal logic to give expression to on-going conditions such as those represented by the progressive aspect in natural language. In this way we are able to enrich more stative mental models so that an agent's internal activities and its perception of external processes can be represented more faithfully. The need for agent consciousness research is briefly discussed.

1 Some 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. Numerous popular and contemporary books by Aleksander [1], Dennett [4], Searle [13], 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. Yet we must admit the possibility that consciousness has utilitarian function, evolved to ensure survival.

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 real 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 although consciousness, in a sentient albeit non-emotive sense, seems more allied to awareness than reasoning, an approximation to human consciousness could enable us to converse more naturally with an individual agent. This makes it an unusual topic of enquiry because we need an account for the first person and second person perspective as well as the more usual third person of objective science.

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 its faithfulness to the human brain. 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 using notions of mental state and practical reasoning 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 [10]. Problematic issues range from neurological syndromes such as phantom limbs, denial of paralysis, and involuntary manipulation, through issues of identity and the effect of emotion and habit, to an explication of context and presence in perception and its links with language. 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, but non-sentient theory like that of [2], which can be realised as a computational agent with a myriad of heterogeneous processes. But to explicate conscious behaviour we certainly require layers of conception which we hardly discuss here.

2 Refining the 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 selecting actions through means-end reasoning in software agents, notably in variants of the Belief, Desire, Intention paradigm. (See, for instance, Bratman[3], Rao and Georgeff [12]). 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. (See, for instance, Labrou and Finin [8]).

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 of *belief*, *desire*, *intention*, *knowledge* and *commitment*, each of which can be regarded intuitively as expressing computational *data* states. Agent activities, or processes, which Vendler (1967) and later workers have considered equally important for the modelling of our linguistic descriptions of behaviour, 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 allow a more refined computational model of rationality. Their absence is a serious deficiency in the usual perception of mental state. However, there is also 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 simply by allowing activity states as well. Both stative and activity states can be considered durative on a temporal frame. They can be distinguished informally by the observation that a stative condition is basically atemporal, but becomes homogeneous on an interval as an artifact of a temporal frame, whereas an activity is essentially durative, a process which may be composed from sub-processes. However, we gloss over finer semantic issues by emphasising one facit, an explicit progressive expression for an activity, captured by a modal operator *prog* to modify a singular predicate for a dynamic

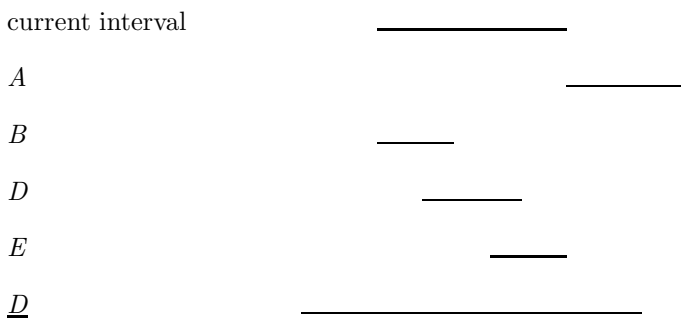

FIG. 1. Interval relations A,B,D,E and $\underline{D}$

verb, so that, for example, a rendering of j *is sensing* c becomes $prog\ senses_j\ c$. To define a *prog* operator we use the interval temporal logic of Halpern and Shoham [7], which uses normal modal operators to incorporate the interval relations identified by Allen. The during relation D, and its complement $\underline{D}$ with respect to the current interval are illustrated in Figure 1, along with basic relations A, B and E (without their complements or other derived operators). Here we simply define *prog* p, or "is p-ing", to be the coercing form $\langle\underline{D}\rangle[D]p$, reading it (right to left) as p *holds on all sub-intervals during some interval which contains the current interval*. Thus this forces homogeneity within some embracing period, but because this serves as an approximation which can be revised to accommodate the substructure of a process. This use of the Halpern and Shoham logic is adapted from a thesis by Leith [9] which provides a tractable linear time account for temporal and aspectual linguistic phenomena. Its use for refined models of intentionality is being explored in joint work with Kamara.

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

$$perceives_j\ p\ \leftrightarrow\ senses_j\ c\ \wedge\ remembers_j\ (c \rightarrow p) \tag{2.1}$$

expresses the idea that sensory perception amounts to an inferential interaction between autonomous sense and memory recall processes on any interval. From the definition of *prog* we can also derive the subtly different aspectual forms in which the current ongoing nature of a process is emphasized:

$$prog\ perceives_j\ p\ \leftrightarrow\ prog\ (senses_j\ c\ \wedge\ remembers_j\ (c \rightarrow p)) \tag{2.2}$$

$$prog\ perceives_j\ p\ \rightarrow\ prog\ senses_j\ c\ \wedge\ prog\ remembers_j\ (c \rightarrow p) \tag{2.3}$$

Although it might be more realistic to consider a compound perception process as composed from particular sense mechanisms, and consider circumstances where the principal implication in the last formula can be reversed, perhaps the least explicable element of these axioms is inferential memory recall itself. Here we have supposed a subconscious associating process rather than an introspective belief state, because as mentioned above this would already be too strong a supposition.

Without any appeal to the notion of consciousness, the use of the progressive in describing a state of affairs enables us to quantify which processes are ongoing in an axiomatic theory of a system, and thus to build an axiomatic theory which can, for example, embrace the detailed effect and management of concurrent processes in refining the intentional stance. The bridge between philosophical notions of mental state and computational systems themselves, is a notion of abstraction which captures externally perceived "emergent" properties of data or process state. It is because consciousness appears to be an abstract process that we address in it this paper.

3 Introspective Awareness

A perception process like that posited by the axioms above could be that of a sophisticated but unconscious automaton. We also consider the awareness needed for consciousness to be a mental process rather than a data state, but if so it is one which includes, in addition to sensory perception, the meta perception of sensory perception. Thus to introduce consciousness, and ultimately a consciousness of responsibility, we need the activity of being aware to be positively introspective. When an agent is aware, it not only perceives, but being sentient, it perceives that it perceives. Thus assuming positive meta perception only, we might provide an axiom for a progressive form of introspective awareness as:

$$aware_j\ p \ \leftrightarrow\ perceives_j\ p\ \wedge\ perceives_j\ perceives_j\ p. \tag{3.1}$$

Again this obscures distinct perception processes, and in so doing may be too strong, 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, some introspective, first person form of perception seems necessary for a sense of identity, so that perception can be become self-awareness. Whether it matters which particular senses that are employed to give this perception, and whether the perception processes are involuntary, or dependent on specific senses is unclear, but something for a deeper investigation along the lines suggested in the next section. A self-referential axiom is certainly no mystery for the implementer of an artificial system. It is indicative of cyclic processing of information, with concommitant issues in bounding any discrete computation, but no less would be expected.

Human 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. Carl Ginet argues for such philosophical processing abstractions and suggests that mental control processes which relate to notions of will and causality can be associated with neurological elements such as the motor cortex. This presumes the ability to effect action, and to control awareness through the ability to switch on perception, or at least to alter the sensitivity to sensory signals. Yet these are merely primitive instances of activities which in more visible forms become motor skills. So it seems that the ability to *will* attention to a selective perception process, or to *will* an action, is a primitive output act for the biological brain. But for this controlling will to be effective the agent, biological or not, must have some capability for the perception or the action, a capability that may have been learned. So we presume a capability before the will can become

causal:

$$capable_j\ p\ \wedge\ wills_j\ p\ \rightarrow\ p. \tag{3.2}$$

Now for an agent to be conscious of a willed activity we suppose it must also be aware of it, and this awareness itself be subject to the will, the willing of a sensory activity, $wills_j\ aware_j\ p$. So we begin to have layers of action and perception in consciousness; something that is willed may also be perceived, and reflexively so. But do we to need senses through which the will itself is perceived? We think not.

The learning process whereby a kitten comes to recognise its tail as its own, and to control it, observations that a familiar activity once started by will can then proceed autonomously and subconsciously until further attention is required, the evidence of breakdown from brain injury of human motor control and to awareness of limb movement, all alert us to the need for an explication of consciousness which encompasses the distinction between voluntary and involuntary action. Yet if we defer consideration of learning itself, and instead consider the invoking of a stored activity plan, and at a greater level of detail the invoking of planned motor actions, we have processes that already fall within the established theories of agent intentionality. All we require are primitive processing actions to will a *beginning* and an *ending* to a durative process, $wills_j\ \langle B\rangle p$, $wills_j\ \langle E\rangle p$, mental actions whereby an agent j causes neural processes for beginning or ending the motor signals for p. Again, the choice of which action to will falls within the deliberative processes of reasoning itself. We can speculate on the degree to which a conscious agent may be aware of these deliberative processes. Learning processes can evidently also be willed, but since there is also involuntary learning, albeit with sophisticated perception involving haptic senses in the aquisition and practise of mental and physical skills, it seems that learning as a process can proceed involuntarily and is not inherently conscious.

To consider an agent conscious is also a social, second and third person judgement of the agent's awareness. So an agent with flawed mechanism for action makes judgement difficult; the agent has lost capability. Yet if internal perception could be observed by experimentation, and if we could observe the ineffective *will* process of a flawed agent, our judgement of whether the agent is conscious would be determined by detecting reaction, and in part social reaction. We expect an intelligent agent that is conscious to display this consciousness in a socially aware way, one which can be distinguished from sleep walking because an intelligent agent which perceives its environment will also perceive and react to social consequences of the relationships in it. Because an agent with more cognitive processing ability could also learn to perceive more complicated, dynamic, causal, social relationships in an environment, and act upon them in deliberated communicative ways, we might also consider such an agent to have heightened consciousness. Indeed, some introspective form of social awareness seems necessary for a sense of responsibility.

So at what stage does consciousness arise? We propose that once an agent has mental activities of a sufficiently introspective perception it also has a form of consciousness, that in its weakest form consciousness is simply a progressive activity of introspective awareness.

$$conscious_j\ \leftrightarrow\ \exists p.prog\ aware_j\ p \tag{3.3}$$

This is a very weak form of sentient consciousness, but graded degrees dependent

on a mental architecture can follow, and perhaps ultimately social consciousness as awareness of social relationships and their consequences for the agent.

4 The Need for Agent Consciousness Research

For the development of artificial communicative agents the need for a clearer picture of social consciousness seems more imperative than the need for biological realism. Nevertheless one expects over time to have productive interplay between the engineering sciences and the sciences of biological organisms. Since the development of effective human-agent communication is of particular interest for the evolution of artificial minds, any correspondence between agent dysfunction and human communicative disorder also becomes of interest. An axiomatic theory, when refined beyond this tenuous start, may also serve as a more succinct and less biased way of describing mental characteristics than explanation using computational or information network models.

If we speculate briefly on a non-social role for biological consciousness as introspective awareness it must still to be concerned with survival fitness. Both biological and non-biological agents have external threats where self awareness and protection is a factor, even though they differ dramatically in the way fitness and mutation are realised and determine evolution. So although the value of a sentient consciousness in this role is less evident, in the design of intelligent agents with limited resources the introspective management of acquired knowledge becomes a factor in the continued usefulness of an agent. Problems of this nature tend to be solved by software designers using knowledge interchange and communication, but an organisationally self-aware agent might add fitness for this purpose and come closer to the non-sentient concept of Barr's consciousness as realised in the "Conscious" Mattie experiment (see Franklin and Graesser [6]). Even if its value is an open issue, a case for exploration of consciousness in agent design is indisputable, whether to better organise complex internal processes, to create a new generation of mutating and evolving self-aware web agents, or merely as defence against such alien creatures.

While there are already workable theories of practical reasoning to guide the construction of reasoning agents, there is also scope for more refined models of intentionality to provide a better way of combining reactive and deliberative responses in a real time agent environment. We have indicated ways in which a sentient form of consciousness can be accommodated within such refinement. While it is possible that what passes for human social skills are learned patterns of behaviour with potentially deliberative purpose, we do not know this. We do know that the planning and understanding of communicative acts as intentions is complex. It needs considerable improvement in the state of the art to allow natural human-agent interaction. So there is certainly scope for more socially sensitive models of interaction where agent identity and the different persons of speech are embraced. While better human-agent interaction may also require treatments of emotion [11], it would even be surprising if a socially competent agent did not also have a socially aware sense of consciousness because an intentional communicative act is planned with a perception of society as context. So elucidation of social consciousness should provide guidelines for future communicative agents.

5 Acknowledgement

The author is grateful for encouragement and suggestions from reviewers, despite the speculative nature of the paper.

References

[1] I. Aleksander, *Impossible Minds: My Neurons, My Consciousness*, Imperial College Press, London, 1996.

[2] B.J. Baars. *In the theater of consciousness*. Oxford University Press, 1997.

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

[4] D. Dennett. *Consciousness Explained*. Penguin, 1991.

[5] C. Ginet. *On Action*. Cambridge University Press, 1990.

[6] S. Franklin and A. Graesser. A Software Agent Model of Consciousness. *Consciousness and Cognition*, **8**, 285–301, 1999.

[7] J. Halpern and Y. Shoham. A Propositional Modal Logic of Time Intervals. In *Proceedings of Symposium on Logic and Computer Science*, Cambridge, MA, 279–292, 1986.

[8] Y. Labrou and T. Finin. Semantics and Conversations for an Agent Communication Language. In M.N. Huhns, and M.P. Singh (Eds.) *Readings in Agents*, 235–242, Morgan Kaufman, San Francisco, CA, 1998.

[9] M.F. Leith. *Modelling Linguistic Events*. PhD Thesis, Imperial College, University of London, 1998.

[10] M. Ito, Y. Miyashita and E. Rolls (eds). *Cognition, Computation, Consciousness*. OUP, 1997.

[11] R.W. Picard, 1997, *Affective Computing* MIT Press

[12] A.S. Rao and M.P. Georgeff. BDI agents:from theory to practice. *ICMAS-95*, San Francisco, 312–319, 1995.

[13] J. Searle. *Mind, Language and Society*. Orion, 1999.

[14] Z. Vendler, *Linguistics in Philosophy*, Cornell University Press, Ithaca, New York, 1967

Received 21 December, 2000