Borrower: BUF

Lending String: *YSM,SYB,RRR,PMC,VXW

Patron: Rapaport, William

Journal Title: Cognition and brain theory.

Volume: 6 **Issue:** 4
Month/Year: 1983**Pages:** 499-508

Article Author:

Article Title: Philip N. Johnson-Laird; Computational Analysis of Consciousness

Imprint: Hillsdale, N.J. ; L. Erlbaum Associates,

ILL Number: 40233514

ARIEL
Charge
Maxcost: $30.00IFM

Shipping Address:
234 Lockwood Memorial Library, (LAND)
University of Buffalo
Interlibrary Loan
Buffalo, NY 14260-2200

Fax:
Ariel: 128.205.111.1

N/C

COGNITION AND BRAIN THEORY, 1983, 6(4), 499-508

Theoretical Notes

A Computational Analysis of Consciousness

P. N. JOHNSON-LAIRD
Medical Research Council
England

ABSTRACT

This paper argues that the problems of consciousness can be solved only by adopting a computational approach and by setting a number of tractable goals for the theory. Four such phenomena need to be explained: the division between conscious and unconscious mental processes; the relative lack of conscious control over many emotions and behaviors; the unique subjective experience of self-awareness; and that aspect of intentionality that is missing from goal-directed computer programs and automata. The paper outlines a theory of consciousness based on three main components: hierarchical parallel processing; the ability to embed models within models recursively; and the possession of a high-level model of the options of the operating system controlling the parallel hierarchy. The first two of these notions are relatively well understood; the third notion is more problematical. However, if the thesis of the paper is correct, the four problems of consciousness can be solved once we understand what it means for a computer program to have a high-level model of its own operations.

THE PROBLEMS OF CONSCIOUSNESS

What should a theory of consciousness explain? This is perhaps the first puzzle about consciousness because unlike, say, the mechanism of inheritance, it is not clear what needs to be accounted for. One might suppose that a putative theory should explain what consciousness is, and how it depends on the workings of the brain. The trouble is that there are no obvious criteria for assessing the success of such a theory. Indeed, this lack of

Requests for reprints should be sent to P. N. Johnson-Laird, Medical Research Council, Applied Psychology Unit, 15 Chaucer Road, Cambridge CB2 2EF, England.

criteria lent respectability to the once popular Behavioristic doctrine that consciousness is not amenable to scientific investigation: it is a myth that the proper study of nerve, muscle, and behavior, will ultimately dispel. A prudent strategy is therefore to take both a different approach to consciousness and to suggest a more tractable set of problems for the theory to solve. My approach here will be to assume that consciousness is a computational matter that depends primarily on the software of the brain rather than on its physical constitution. And there are, I believe, four principal problems that any theory of consciousness should solve:

1. *The problem of awareness.* When someone speaks to you, you can be aware of the words they utter, you can be aware of the meaning of their remarks, and you can be aware of understanding what they are saying. And, if none of this information is available to you, then you can be aware of that, too. Yet, there is much that is permanently unavailable to you. You cannot be aware of how you understand the speaker's remarks or of the form in which their meaning is represented in your mind. In fact, you can never be completely conscious of how you exercise any mental skill. A theory of consciousness must account for this division between conscious and unconscious processes; it is a striking clue to the architecture of the mind, but it is so familiar to us, and so easily confused with distracting Freudian notions, that its importance has been overlooked.

2. *The problem of control.* You cannot control all of your feelings. You can feign happiness and sadness, but it is difficult, if not impossible, for you to evoke these emotions merely by a conscious decision. Conversely, a particular feeling may overwhelm you despite all your efforts to resist it. This lack of control extends, of course, to behavior. You may, for instance, have a genuine intention to give up smoking but be unable to put it into practice. Some individuals can exert a tight control on themselves and on their expressions of emotion; others, like Oscar Wilde, can resist everything except temptation. There do indeed seem to be differences in will power from one individual to another, though the topic is almost taboo in cognitive psychology.

3. *The problem of self-awareness.* When you are doing something like driving a car, you can be aware of what you are doing or you can be so absorbed as to forget yourself. When you are aware of what you are doing, you may become aware of the fact that you are aware of what you are doing. Sometimes, perhaps, you can even be aware of the fact that you are aware that you are aware, and so on. The ability to be aware of what one is doing is, of course, essential for self-awareness and for one's sense of the integrity, continuity, and individuality of the self.

4. *The problem of intentionality.* To act intentionally is at the very least to decide to do something (for some reason or to achieve some goal) and

then in conseq
that are gover
ming language
and that then
deterministic
they had inte
them because
consciousness
ing from such

Psychologi
sciousness. T
more compl
culminating
1934). But
neurones wit
simple acts;
sciousness,
our slumber
with the co
Boies, 1971)
goals to see
processing
1975). Thes
lems above
computer r
putational
solved prob

From the
human bra
chronous
is also bor
ferent lev
processed
suppose th
internal w
physically
sequences
municatio

ic doctrine that
: is a myth that
nately dispel. A
proach to con-
or the theory to
ss is a computa-
he brain rather
, four principal

you, you can be
ıeaning of their
they are saying.
ou can be aware
ıble to you. You
marks or of the
In fact, you can
mental skill. A
tween conscious
:hitecture of the
with distracting
d.
ur feelings. You
: impossible, for
n. Conversely, a
forts to resist it.
ay, for instance,
ole to put it into
emselves and on
resist everything
es in will power
almost taboo in

g something like
you can be so ab-
it you are doing,
what you are do-
fact that you are
re of what one is
's sense of the in-

s at the very least
e some goal) and

then in consequence to do it. There are many computer programs, however, that are governed by internally generated goals, for example, the programming language PLANNER enables programs to be written that set up goals, and that then seek to achieve those goals by simulating the action of a non-deterministic automaton (see Hewitt, 1972). These programs act as though they had intentions, but it would be a mistake to ascribe intentionality to them because they have no awareness of what they are doing. A theory of consciousness should elucidate the component of intentionality that is missing from such programs.

Psychologists and others have, of course, proposed theories of consciousness. They have tried to account for it in terms of the evolution of more complex brains (e.g. John, 1976), or more complex behaviors culminating in linguistic communication and social relations (e.g. Mead, 1934). But consciousness is hardly a necessary consequence of more neurones with more connections between them; we can be conscious of very simple acts; and, if language and society could have evolved without consciousness, why should they need, and how would they be able to awaken our slumbering minds? Psychologists have also identified consciousness with the contents of a limited capacity processing mechanism (Posner & Boies, 1971), with a device that determines what actions to take and what goals to seek (Shallice, 1972), and with a particular mode of information processing that affects the mental structures governing actions (Mandler, 1975). These claims are plausible, but they do not solve all of the four problems above, and they might even be taken to apply to a device such as a computer running a PLANNER-like program. My aim is to sketch a computational approach to consciousness that may lead to solutions to the unsolved problems of awareness, control, self-awareness, and intentionality.

HIERARCHICAL PARALLEL PROCESSING

From the simple nerve networks of coelenterates to the intricacies of the human brain, there appears to be a uniform computational principle: asynchronous parallel processing. That higher mental processes occur in parallel is also borne out by the fact that, for example, language is organized at different levels—speech sounds, morphemes, sentences, and discourse—and processed at these levels contemporaneously. There are good reasons to suppose that one processor in the parallel system cannot directly modify the internal workings of another, because such interactions—even if they were physically possible—would produce highly unstable and unpredictable consequences. A more plausible form of interaction relies solely on the communication of messages between the processors. These messages may take the form of predictions, constraints on processing, the results of computations, emergency signals, and other such interrupts. The most general conception of a system of parallel processing meeting this communicative constraint is of a set of finite-state automata with channels between them for communicating the values of parameters and global variables, that operate according to the principle that each processor starts to compute as soon as it receives the values of the parameters and variables that it needs. Other parallel systems are special cases of this design, for example, systems in which only information about level of activation is passed from one processor to another (see Anderson & Hinton, 1981), systems in which all the processors are synchronized by reference to an internal clock (see Kung, 1980), and vector machines in which all the processors carry out the same procedure (see Kozdrowicki & Theis, 1980). It is important to keep in mind the distinction between a function and an algorithm for computing that function, because there are infinitely many different algorithms for computing any computable function (see e.g. Rogers, 1967). Moreover, although any function that can be computed in parallel can be computed by a serial device, there are many algorithms that run on parallel computers that cannot run on serial ones. Hence, if consciousness depends on the computations of the nervous system, then it is likely to be a property of the algorithms that are used to carry out those computations rather than a property of their results (Johnson-Laird, 1983): it ain't what you do, it is the way that you do it!

There are some problems—the parsing of certain abstract languages, for example—that can be shown to be solvable in principle but not in practice: they are to computation what Malthus's doctrine of population growth is to civilization. A problem is inherently intractable when any algorithm for it takes a time that grows exponentially with the size of the input (see e.g., Hopcroft & Ullman, 1979). For an input of *n*, where *n* is small, an algorithm may be feasible, but even if the time it takes is proportional, say, to 2^n, then, because such exponentials increase at so great a rate with an increase of *n*, a computer the size of the universe operating at the speed of light would take billions of years to compute an output for a relatively modest input. Parallel processing is of no avail for rendering such problems tractable. What it does is to speed up the execution of algorithms that take only a time proportional to a polynomial of the size of the input. If many processors compute in parallel, they can divide up the task between them whenever there are no dependencies between the computations. Such a division of labor not only speeds up performance, but it also allows several processors to perform the same sub-task so that should one of them fail the effects will not be disastrous, and it enables separate groups of processors to specialize in different sub-tasks. The resulting speed, reliability, and specialization have obvious evolutionary advantages. But parallel computation has its dangers too, for example, if the radial nervous system of a starfish is divided into two separate arcs, the organism may tear itself apart in trying to move in opposite directions. Any simple nervous system with a fixed program that behaved in this way would soon be eliminated by natural selection. Unfortunately, higher organisms do not appear to have fixed software—they can implement new programs to meet unexpected contingencies—and there must therefore be mechanisms, other than those of direct selective pressure, to deal with processing conflicts. A sensible design here is to promote one processor to monitor the operations of others and to override them in the event of conflicts and other, more deadly, pathological states of affairs. If this design feature is replicated on a large scale, the resulting architecture is a hierarchical system of parallel processors: a high-level processor that controls the overall goals of lower-level processors, which in turn monitor the processors at a still lower level, and so on down to the lowest level of processors that govern sensory and motor interactions with the world. A hierarchical organization of the nervous system has indeed been urged by neuroscientists from Hughlings Jackson to H. J. Jerison (see Oatley, 1978, for the history of this idea), and Simon (1969) has argued independently that it is an essential feature of intelligent organisms.

THE OPERATING SYSTEM

In a hierarchical computational device, the highest level of processing consists of an operating system. The operating system of a digital computer is a suite of programs that allows a human operator to control the computer. There are instructions that enable the operator to recover a program stored on a magnetic disk, to compile it, to run it, to print out its source code, and so on. When the computer is switched on, its resident monitor is arranged to load the operating system either automatically or as a result of some simple instructions. The notion that the mind has an operating system verges, as we shall see, on the paradoxical, but it has some relatively straightforward consequences. The operating system must have considerable autonomy, though it must also be responsive to demands from other processors. It must be switched on and off by the mechanisms controlling sleep. It must depend on a second level of processors for perceiving, understanding, acting, remembering, communicating, and thinking. These processors in turn must depend on lower level processors for passing down more detailed control instructions or for passing up partially interpreted sensory information. Doubtless, there are interactions between processors at the same or different levels, and facilities that allow priority messages from a lower level to interrupt computations at a higher level. The hierarchy of communicating parallel processors imposes one great virtue on the operating system: it can be relatively simple, because it does not need to be concerned with the detailed implementation of the instructions that it sends to lower level processors. It specifies what they have to do (e.g., to walk, to think, to talk) but not how they are to carry out the computations that underlie these tasks. It receives information from the lower level processors about the results of computations, but not about how they were obtained. Thus, the visual world is presented to us in a way that is as real as the stone that Dr. Johnson kicked in order to refute Idealism: we have no access to the sequence of representations that vision must depend on (Marr, 1981). The phenomenal reality of the visual world is a triumph of the adaptive nature of the mind. If we were aware that the visual world is a representation, then we would be more likely to doubt its veridicality and to treat it as something to be pondered over—a potentially fatal debility in the event of danger. Psychology is difficult just because there is an evolutionary advantage in a seemingly direct contact with the world and in hiding the cognitive machinery from consciousness.

SOME EMPIRICAL EVIDENCE

Let us take stock of the argument so far. The brain is a parallel computer that is organized hierarchically. Its operating system corresponds to consciousness and it receives only the results of the computations of the rest of the system. Such a system can readily account for the division between conscious and unconscious processes (the problem of awareness) and it can also allow the lower level processors a degree of autonomy (the problem of control). There are at least three clinical syndromes that corroborate this division. First, there is the phenomenon of "blind-sight" described by Weiskrantz, Warrington, Sanders, and Marshall (1974). After damage to the visual cortex, certain patients report that they are blind in parts of the visual field, and their blindness is apparently confirmed by clinical tests. Yet, more subtle testing shows that the patients are able to use information from the 'blind' part of the field. It seems that their sight in the affected regions has continued to function but no longer yields an output to the operating system: they see without being conscious of what they see. Second, there are the 'automatisms' that occur after epileptic attacks. In this state, patients seem to function completely without consciousness and without the ability to make high-level decisions. They may be capable of driving a car, for example, but unable to respond correctly to traffic lights (Penfield, 1975). Evidently, the attack leads to a dissociation between the operating system and the multiple processors. Third, there are the well-attested cases of hysterical paralysis. Prolonged stress may lead to paralyses that have, unbeknownst to the patient, no neurological explana-

tion. They ca
World War I,
stimuli will p
tients are not
that is outsid

SELF-A

Granted the
parallel hiera
phenomenal
two forms o
jective feelir
proach so fa
might be fou
other proces
operating sy
that the hur
system actu
right to the

Reflectio
the followi
to some ext
The idea is
as paradox
England we
then it sho
map depict
tion of itse
on ad infir
cannot be
of the min
putational
itself recu
diminishin
of the vari
would go

There is
understan
embeddin
ubiquitou
tents of o
ple's min

tion. They can often be cured, as the late Lord Adrian showed during World War I, by similarly duping the patient into believing that electrical stimuli will produce a cure (see Adrian & Yealland, 1917). Since these patients are not malingering, they provide us with clear examples of a reaction that is outside the knowledge and control of the operating system.

SELF-AWARENESS AND THE EMBEDDING OF MODELS

Granted the hypothesis of a serial operating system communicating with a parallel hierarchy, one might ask how the former gives rise to the singular phenomenal experience of consciousness, and why there couldn't just be two forms of computation carried out by the brain with no particular subjective feeling associated with either. Evidently, the computational approach so far accounts only for a bare awareness of the world, such as might be found in other species, for the division between that awareness and other processes, and for the phenomena of control. Moreover, a computer's operating system does not make decisions, but merely implements those that the human operator conveys to it. How could the mind's operating system actually make conscious decisions? This question, of course, goes right to the core of the problems of self-awareness and intentionality.

Reflection on the human capacity for self-reflection leads inevitably to the following observation: the mind is aware of itself. It understands itself to some extent, and it understands that it understands itself, and so on. . . . The idea is central to the subjective experience of consciousness, yet it seems as paradoxical as the conundrum of an inclusive map. (If a large map of England were traced out in accurate detail in the middle of Salisbury plain, then it should contain a representation of itself within the portion of the map depicting Salisbury plain (which in turn should contain a representation of itself (which in turn should contain a representation of itself (and so on ad infinitum)))). Such a map is impossible because an infinite regress cannot be captured in a physical object. Leibniz dismissed Locke's theory of the mind because there was just such a regress within it. However, a computational procedure for representing a map can easily be contrived to call itself recursively and thus to go on drawing the map within itself on an ever diminishing scale. The procedure could in principle run for ever: the values of the variables, though too small to be physically represented in a drawing, would go on diminishing perpetually.

There is a similar computational solution to the paradox of the mind understanding the mind. Human beings have the ability to make recursive embeddings of mental models within mental models. Such embeddings are ubiquitous in cognition. For example, we all construct models of the contents of other people's minds, and sometimes of their models of other people's minds, and so on. The recursive aspect of this ability is hardly problematical. The crux of the problem of consciousness resides in another requirement: the operating system must have a partial model of itself in order to begin the process of recursion underlying self-awareness.

No one knows what it means to say that an automaton or computer program has a model of itself. The question has seldom been raised and certainly has yet to be answered. The notion must not be confused with self-description (pace, Minsky, 1968). It is a relatively straightforward matter to devise an automaton that can print out its own description (see e.g., Thatcher, 1963). But such an automaton merely advertizes its own inner structure in a way that is useful for self-reproduction, and it no more understands that description than a molecule of DNA understands genetics. What is needed is a program that has a model of its own high-level capabilities. This model would be necessarily incomplete, according to the present theory, and it might also be slightly inaccurate, but it would nonetheless be extremely useful. People do indeed know much about their own high-level capabilities: their capacity to perceive, remember, and act; their mastery of this or that intellectual or physical skill; their imaginative and ratiocinative abilities. They obviously have access only to an incomplete model, which contains no information about the inner workings of the web of parallel processors. It is a model of the major options available to the operating system.

CONCLUSIONS

A complete theory of consciousness depends on putting together the three main components that I have outlined: hierarchical parallel processing, the recursive embedding of models, and the high level model of the system itself. Self-awareness and intentionality depend on a recursive embedding of a model of the self within itself so that the different embeddings are accessible in parallel to the operating system. To have a conscious intention, for instance, the operating system must elicit a representation of a possible state of affairs, and decide that it itself should act so as to try to bring about that state of affairs. An essential part of this process is precisely an awareness that the system itself is able to make such decisions. The system has to be able to represent the fact that the system can generate a representation of a state of affairs and decide to work towards bringing it about. At a low level, there is a program (perhaps analogous to a program in PLANNER) that can construct a model of a state of affairs, and act so as to try to achieve it. But the system can construct a model of itself operating at this low level of performance, and it can use this model in the process of making a decision. It can also construct a model of its own performance at this higher level in turn, and so on . . . to any required degree of embedding. Since the hier-

archy of embe
make consciou
to be aware o

The present
putations of t
voked are, wi
itself, reasona
ly the latest i
engines, telep
the brain. Wl
ceeded in refu
computable.
functions the
the last meta

Adrian, E. D.,
1917, June, 3
Anderson, J. A
Hinton and J
Lawrence Erl
Hewitt, C. *Des*
dum AI TR-2
Hopcroft, J. E.
Mass.: Addis
John, E. R. A
ness and self
Johnson-Laird,
Mass.: Harva
Kozdrowicki, E
1980, *13*, 71-
Kung, H. T. T
puters, Vol.
Mandler, G. A
Marr, D. *Visio*
mation. San
Mead, G. H.
C. W. Morr
Minsky, M. L.
essing. Cam
Oatley, K. *Pe*
chology. Lo
Penfield, W.
Posner, M. I.
391–408.

archy of embedded models exists in parallel, the operating system is able to make conscious decisions, to be aware of making conscious decisions, and to be aware of its own awareness.

The present approach assumes that human behavior depends on the computations of the nervous system. The class of procedures that I have invoked are, with the exception of a program that has a high-level model of itself, reasonably well understood. It is often said that the computer is merely the latest in a long line of inventions—wax tablets, clockwork, steam engines, telephone switchboards—that have been taken as metaphors for the brain. What is often overlooked, however, is that no one has yet succeeded in refuting the thesis that any explicit description of an algorithm is computable. If that thesis is true, then all that needs to be discovered is what functions the brain computes and how it computes them. The computer is the last metaphor for the mind.

REFERENCES

Adrian, E. D., & Yealland, L. R. The treatment of some common war neuroses. *Lancet*, 1917, June, 3–24.

Anderson, J. A., & Hinton, G. E. Models of information processing in the brain. In G. E. Hinton and J. A. Anderson (Eds.), *Parallel models of associative memory.* Hillsdale, N.J.: Lawrence Erlbaum Associates, 1981.

Hewitt, C. *Description and theoretical analysis (using schemata) of PLANNER.* Memorandum AI TR-258. MIT Artificial Intelligence Laboratory, 1972.

Hopcroft, J. E., & Ullman, J. D. *Formal languages and their relation to automata.* Reading, Mass.: Addison-Wesley, 1979.

John, E. R. A model of consciousness. In G. E. Schwartz and D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research, Vol. 1.* London: Wiley, 1976.

Johnson-Laird, P. N. *Mental models.* Cambridge: Cambridge University Press. Cambridge, Mass.: Harvard University Press, 1983.

Kozdrowicki, E. W., & Theis, D. J. Second generation of vector super-computers. *Computer*, 1980, *13*, 71–83.

Kung, H. T. The structure of parallel algorithms. In M. C. Yovits (Ed.), *Advances in computers,* Vol. 19. New York: Academic Press, 1980.

Mandler, G. *Mind and emotion.* New York: Wiley, 1975.

Marr, D. *Vision: A computational investigation in the human representation of visual information.* San Francisco: Freeman, 1982.

Mead, G. H. Mind, self and society from the standpoint of a social behaviorist. Edited by C. W. Morris. Chicago: University of Chicago Press, 1934.

Minsky, M. L. Matter, mind, and models. In M. L. Minsky (Ed.), *Semantic information processing.* Cambridge, Mass.: MIT Press, 1968.

Oatley, K. *Perceptions and representations: The theoretical bases of brain research and psychology.* London: Methuen, 1978.

Penfield, W. *The mystery of mind.* Princeton: Princeton University Press, 1975.

Posner, M. I., & Boies, S. J. Components of attention. *Psychological Review*, 1971, *78*, 391–408.

Rogers, H. *Theory of recursive functions and effective computability.* New York: McGraw-Hill, 1967.

Shallice, T. Dual functions of consciousness. *Psychological Review,* 1972, *79,* 383-393.

Simon, H. A. *The sciences of the artificial.* Cambridge, Mass.: MIT Press, 1969.

Thatcher, J. W. The construction of a self-describing Turing machine. In J. Fox (Ed.), *Mathematical theory of automata.* Microwave Research Institute Symposia, Vol. 12. Polytechnic Institute of Brooklyn, New York: Polytechnic Press, 1963.

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain,* 1974, *97,* 709-728.

Critical D

Ph

Husserl is be
find in his w
model of mi
advocate of
tion of essay
argues that
similar to Je
bodies all th
Many seem t
of the enthu
he solved sor
I want to tal
Husserl's im

Reading H
ambitious pl
he actually a
sometimes r
ties through
readers to
Husserl con
of mind or

There are
Husserl was
"scientifical
systematical
discovers in

Requests for
Clung Tower,