Brain Inspired Cognitive Systems
August 29 – September 1, 2004
University of Stirling, Scotland, UK

MECHANICAL BODIES; MYTHICAL MINDS

Dr. Mark Bishop, Dept. Computing, Goldsmiths College, New Cross, London, UK

ABSTRACT

A cursory examination of the history of Artificial Intelligence, AI, serves to highlight several strong claims from its researchers, especially in relation to the populist form of computationalism that holds, 'any suitably programmed computer will instantiate genuine conscious mental states purely in virtue of carrying out a specific series of computations'.

The argument to be presented in this paper develops ideas first outlined in Hilary Putnam's 1988 monograph, "Representation & Reality", then developed by the author in, "Dancing with Pixies", (2002a) and "Counterfactuals Cannot Count", (2002b). This work further extends these ideas into a novel thesis against computationalism which, if correct, has important implications for Cognitive Science; both with respect to the prospect of ever developing a computationally instantiated consciousness and more generally for any computational, (purely-functional), explanation of mind.

INTRODUCTION

Many people find the notion of machine phenomenology so difficult to accept that, in this aspect at least, arguments against computational consciousness target a straw man. Yet a quick search of AI literature reveals many eminent cognitive scientists, including Minsky, (1985), Moravec, (1998) and Kurzweil, (1998), who have speculated positively on the subject; with some of the 'new-roboticists', Warwick (1996), O'Regan & Noe, (2001a, 2001b) and Harvey, (2002), specifically claiming that if devices exhibit appropriate sensori-motor co-ordination then there is no 'in-principle' barrier to conceiving of them as conscious devices and indeed that, we have already developed robots '*as conscious as a slug*' (Warwick, 2002). It is to this group that the following reductio is directed.

The main argument presented here is not significantly original – it is a simple reflection upon that originally given by Hilary Putnam (Putnam 1988) and widely criticised by David Chalmers and others[1]. However, in what follows, instead of seeking to justify Putnam's original claim that, "every open system implements every finite state automaton", (FSA), and hence that psychological states of the brain cannot be functional states of a computer, I will seek to establish the weaker result that, over a finite time window every open system implements the trace of a particular FSA Q, as it executes with known input (x). That this result leads to panpsychism is clear as, equating Q (x) to a specified program that is claimed to instantiate phenomenal states as it executes, and following Putnam's procedure, identical computational (and *ex-hypothesi* phenomenal) states can be found in every open physical system.

The route-map for this endeavour is as follows. In the first part of the paper I review the *Dancing with Pixies* reductio ad absurdum argument, (Bishop, 2002a), against computationally instantiated conscious states; then I review several responses to the argument and conclude by showing how, if the reductio holds, it undermines not just the notion of machine consciousness but more generally any computational explanation of mind.

BACKGROUND

In recent years the most well known arguments against computational explanations of mind have come from John Searle, in the Chinese Room, (Searle, 1980) and Roger Penrose's application of the Godelian argument to the analysis of how, "mathematicians in general provide their 'unassailable demonstrations' of the truth of certain mathematical assertions", (Penrose, 1989 & 2002). However, less well known outside the field, is Hilary Putnam's conclusion that, "if true, functionalism implies behaviourism", published as an appendix to his 1988 monograph, "Representation & Reality"[2].

Central to Putnam's conclusion is his proof of the theorem, "*Every ordinary open system is a realization of every abstract*

[1] See Chalmers (1994, 1996a, 1996b) and also the special issue, *What is Computation?* of Minds and Machines, (vol.4, no.4, November 1994).

[2] Ironically Putnam is widely considered to be the father of Functionalism as a philosophical theory of Mind.

finite automaton". This theorem has attracted much debate, with a special edition, (1994, 4:4), of the journal 'Minds & Machines' devoted to teasing out the notion of exactly 'What is computation?' Subsequently, in a Synthese article published in 1996(b), David Chalmers seemed to fatally undermine Putnam's attack on functionalism by demonstrating that, if an open system is to fully realise even the simplest computation with input, "in a very short time, the system will be larger than the known universe".

DANCING WITH PIXIES, (DWP)

In 2002 the author's paper, *Dancing with Pixies*, sought to obviate Chalmers response to Putnam's theorem with respect to any claimed phenomenal states of a putative conscious computer/robot. The essence of DWP is the following reductio ad absurdum:

1. If it, (our assumed claim), is true, "***that an appropriately programmed computer really has genuine cognitive states***"

2. Then "***panpsychism holds***".

3. However, against the backdrop of our immense scientific knowledge of the physical world, and the corresponding widespread desire to explain everything ultimately in physical terms, panpsychism has come to seem an implausible view.

Hence we should reject the assumed claim (1).

Clearly the core work of the reductio is to establish the implication linking steps [1] and [2]. This is achieved by a tightly constrained application of Putnam's theorem. I.e. We do not seek to establish that, "*every ordinary open system is a realization of every abstract finite automaton*", simply the weaker result that, "*over a finite time period, everything implements the trace of Finite State Automata Q as it operates on fixed input (x)*".

In DWP it is shown that this weaker result does not lead to the combinatorial explosion in required physical states that Chalmers demonstrated a complete implementation of an FSA with input would require and hence that Putnam's theorem, mapping logical states of a computation to physical states of any open system, holds and panpsychism is true.

THE NEXUS OF PUTNAM'S THEOREM

1. The computational states of a system are always relative to the observed function **and** the underlying physics of the system. i.e. Unlike say mass or form, computational states are not intrinsic to physical states of matter but always require a mapping from physical state to logical state.

2. Domain A: The phenomenal states of a putative conscious computational system are independent of the underlying computational hardware; specific phenomenal state sequences are instantiated by specific sequences of computational modal state transitions.

3. Domain B: The behaviour of any open physical system can be described by a series of modal state transitions.

4. Over a finite bounded interval there exists a simple reliable mapping between these two domains. I.e. In line with any computational system there is mapping from the hardware [the open physical system] and the computational states generated as the program executes:- **the Putnam mapping**.

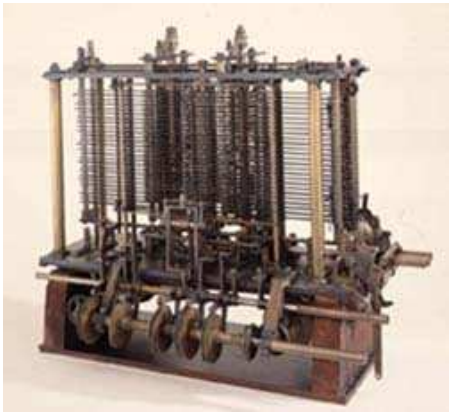
5. Hence if a suitably programmed computer genuinely experiences phenomenal states as the program executes then so does any open system; **Panpsychism is true**.

COMPUTATION IS NOT INTRINSIC TO PHYSICS

In any computer logical states are always mapped onto physical states of the system. For example in a computer built using TTL (Transistor-Transistor Logic), the convention is that 0V maps to logic FALSE and +5V maps to TRUE. In the construction of computers this has mapping not always been the one used; other logic systems, (e.g. RS232), have represented logic TRUE is by a voltage between -15V to -3V and a logic FALSE by +3V to +15V.

COMPUTATIONAL STATES ARE NOT INTRINSICALLY ELECTRICAL

It is exactly the observer-relative mapping of computational states onto physical states that allows different modalities of computation; computers do not have to be electrical. For example Babbage's Difference & Analytical Engines were mechanical devices that mapped logical states onto mechanical states, (see figure below).



A portion of Charles Babbage's Analytic Engine, (Science Museum London)

… as did Weizenbaum's toilet roll and pebble 'machine' that played a simple game, (Weizenbaum 1976, pp.51ff). More recently extremely fast computing systems are being developed that use light instead of electricity to represent logical states.

INPUTLESS FINITE STATE AUTOMATA, (FSA)

An input-less FSA is specified by a set of formal states $\{S_1, S_2, .. S_n\}$, and by a set of state-transition relations which specify for each state the next state that must follow. Given the initial condition(s) an input-less FSA will transit a series of states before eventually entering a cyclic terminal state sequence of one (or more) states.

HOW TO IMPLEMENT ANY INPUTLESS THREE STATE FINITE STATE AUTOMATA WITH A SIMPLE COUNTER

Fundamentally, a system can be said to implement an input-less FSA over a given time-period, if there is a mapping *f* from physical states of the system to computational states of the FSA such that: if the system is in physical state *p* during the time-period, this causes it to transit into a state *q* such that computational state *f (p)* transits to computational state *f (q)* in the specification of the FSA.

Over the time interval $[T_1$ to $T_6]$ a simple digital counter transits the states, $\{C_1, C_2, C_3, C_4, C_5, C_6\}$. Over the same time interval an input-less FSA Q generates the finite linear series of state transitions labelled, $\{Q_1, Q_2, Q_3, Q_1, Q_2, Q_3\}$. Hence to implement the input-less FSA Q by the counter we need to use the following mapping, *f*, from counter states to computational states:

- Map FSA state Q_1 to the disjunction (C_1 v C_4).
- Map FSA state Q_2 to the disjunction (C_2 v C_5).
- … and FSA state Q_3 to (C_3 v C_6).

As is usual for any computational system the mapping simply assigns a logical, computational state onto a physical state of the system. However, by adopting this mapping any simple digital counter will generate the required state transition sequence of our input-less FSA, $\{Q_1, Q_2, Q_3, Q_1, Q_2, Q_3\}$, over the specified time interval.

Note, after Chalmers, that the above counting system will only implement one particular path through the FSA state structure – there may be other state transition sequences that have not emerged in this execution trace. To circumvent this problem Chalmers suggests using a [counter] system with an extra dial – a sub-system with an arbitrary number of states, $[C_{[\text{dial-state, counter-state}]}]$.

Now, as Chalmers suggests, we associate dial-state [1] with the first run of the FSA. The initial state will then be $[C_{[1, 1]}]$ and we associate this with an initial state of the FSA. We then associate system states $[C_{[1, 2]}]$, $[C_{[1, 3]}]$ with associated FSA states using the Putnam mapping described earlier. If at the end of this process some FSA states have not come up, we choose a new FSA state, Q', increment the dial to position [2] and associate this new state $[C_{[2, 1]}]$ with Q' and proceed as before. By repeating this process all of the states of the FSA will eventually be exhausted. Then, for each state of the FSA there will be a non-empty set of associated counter system states. To obtain the FSA implementation mapping we use Putnam's mapping once more and the disjunction of these states is mapped to the FSA state as before. Chalmers remarks:

> "*It is easy to see that this system satisfies all the strong conditionals in the strengthened definition of implementation. For every state of the FSA, if the system is (or were to be) in a state that maps onto that formal state, the system will (or would) transit into a state that maps onto the appropriate succeeding formal state. So the result is demonstrated*", (Chalmers 1996a, p.317).

THE PHYSICAL STATE OF THINGS

Physics typically describes the time evolution of a complex system via a set of dynamic equations. By selecting appropriate intervals a system's behaviour can be quantised into a series of modal state transitions between regions of phase space. Hence a physical system can be characterised by a series of discrete states that evolve over time.

E.g. A simple three state characterisation of water heating in a kettle may be (1) that the water goes from a cold state to a warm state; (2) to a hot state; (3) to a boiling state, (and perhaps eventually to a 'cup-of-tea' state).

Due to influence of cosmic rays, gravitational fields etc. any *Open Physical System* is characterised by a series of non repeating states that evolve over time, $\{S_1, S_2, S_3, S_4, S_5, S_6 .. S_\infty\}$. This non-cyclic behaviour is analogous to that exhibited by an infinite counter, (i.e. one that never repeats states).

THE TEA GOD

Consider that an open physical system, (eg, a cup of tea), over the period $[T_1 .. T_n]$ is described by the physical state transitions $\{S_1, S_2, .. S_n\}$. With correct knowledge of initial conditions and system boundary conditions, a Laplacian Supermind, (The Tea God), can reliably map from system state S at T to S' at T' to S'' at T''. i.e. Given an initial state of the system and the boundary conditions that pertain, The Tea God can reliably predict the future state of the tea at any time.

THEOREM: 'ANY OPEN PHYSICAL SYSTEM IMPLEMENTS ANY INPUTLESS FSA'

Over the time interval $[T_1$ to $T_6]$ an input-less FSA Q generates the finite linear series of state transitions labelled, $\{Q_1, Q_2, Q_3, Q_1, Q_2, Q_3\}$. Any open physical system, (e.g. a cup of tea), transits system states, $\{S_1, S_2, S_3, S_4, S_5, S_6\}$, in same time period.

To implement any input-less FSA Q by an open physical system:

- Map FSA state Q_1 to the disjunction (S_1 v S_4).
- Map FSA state Q_2 to the disjunction (S_2 v S_5).
- … and FSA state Q_3 to (S_3 v S_6).

Once again we add a dial, (perhaps implemented as a scratch on the cup), to ensure all possible traces through the FSA are implemented; now, as in any computational system, the Putnam mapping simply maps a logical, computational state, from the physical state of the system.

By using this mapping any open physical system will generate the required state transition sequence, {Q_1, Q_2, Q_3, Q_1, Q_2, Q_3}, over the specified time interval[3].

Chalmers remains unfazed at this result because he states that inputless FSA's are simply an "inappropriate formalism" for a computationalist theory about the mind:

> "*To see the triviality, note that the state-space of an inputless FSA will consist of a single unbranching sequence of states ending in a cycle, or at best in a finite number of such sequences. The latter possibility arises if there is no state from which every state is reachable. It is possible that the various sequences will join at some point, but this is as far as the 'structure' of the state-space goes. This is a completely uninteresting kind of structure*", (ibid., p.318).

CHALMERS: COMPUTING WITH INPUT AND OUTPUT

The behaviour, (state evolution), of a Finite State Automaton with input, output and memory, (an abstract model of a modern digital computer), is specified by a complex tree of potential input, memory and computational state contingencies. Chalmers demonstrated, (ibid), that to fully implement this form of contingent state structure with an open physical system and a Putnam style mapping an exponential number of system states, (as a function of elapsed time), are required. Hence, as run-time increments, this value rapidly becomes larger than the number of atoms in the known universe and functionalism is preserved…

COMPUTING LIKE CLOCKWORK

Because, in any real computer system memory is finite, the memory state and computational state can be conjoined to form a finite set of 'super-states' of the automata. Further, full knowledge of the input will collapse the complex branching state structure of the automaton to a simple linear path.

E.g. If we are in super-state {A} and the state transition rules specify that: if (input = 'b') we enter super-state {B}; (input = 'c') we enter super-state {C}; if (input = 'd') we enter super-state {D} and if we know the input is defined as ('b'), then we can replace this contingent branching structure by a simple state transition {A} ⇒ {B}.

Further, over any finite interval, all circular (iterative) state transition paths can be unfolded to produce a finite linear series of state transitions.

E.g. Consider the loop cycling through states {Q_1, Q_2, Q_3,} over nine time steps. The iterative loop structure can simply be replaced by the linear series of state transitions, {Q_1, Q_2, Q_3, Q_1, Q_2, Q_3, Q_1, Q_2, Q_3}.

Hence, with its input fixed over a finite time period, an automaton with finite memory simply functions 'like clockwork'.

[3] See Chalmers (1996a) for discussion of reliability and initial conditions.

MECHANICAL BODIES - HAPPY MINDS?

Can a machine, (a robot), be happy? Can a machine experience genuine phenomenal states purely in virtue of executing an appropriate program?

Let the robot's behaviour, given input (x), be defined by the Finite State Automata Q (x). Consider the action of Q (x) over the interval [T_1 .. T_n]. As input is fixed (x), the state transition diagram can be unfolded to a linear path. i.e. Q (x) will generate a finite linear series of state transitions at clock intervals of Q, {Q_1, Q_2, Q_3 .. Q_n}. It is the claim of Artificial Consciousness Researchers that during this time interval, as the robot FSA executes, genuine phenomenal states, (e.g. happiness), are 'mechanically' realised by the machine.

HAPPY TEA?

Is a cup of tea happy? Can a 'cup of tea' experience genuine phenomenal states purely in virtue of traversing a specific series of modal state transitions?

Clearly over any specified time interval [T_1 .. T_n] we can map 'cup of tea' states {S_1, S_2, .. S_n} to robot FSA states {Q_1, Q_2 .. Q_n}, using the Putnam transform.

Just as for the robot FSA, the 'cup of tea' state transitions are modal, (the state transitions are forced; given initial conditions {S_1}, boundary conditions and elapsed time, we can predict any future state of the tea, {S_n}).

Hence, if the claims of Artificial Consciousness Researchers are true, (and the robot experiences phenomenal states - e.g. happiness - purely in virtue of its transit through an appropriate sequence of modal state transitions), then so does a 'cup of tea' and disembodied consciousness lurks in every open physical system; little pixies are dancing everywhere…

OBJECTION 1: HOFSTADTER, "THIS IS NOT SCIENCE"

It has been claimed that such a-posteriori mappings do not qualify as genuine mappings as we can only perform them once we know the input(s) to the robot over the specified time interval. Reflecting on Searle's use of a similar a-posteriori device Doug Hofstadter (1981), once famously declared, "this is not science!"

However consider two identical experiments, (exptA & exptB), performed one after the other for the same period of time on the same 'conscious' robot starting from the same initial state and given input (x_A) and (x_B) where ($x_B = x_A$).

In both exptA and exptB, as the input and initial conditions are identical and the robot is fully deterministic, it must execute an identical series of computational state transitions. Hence there is no principled reason why a-priori knowledge of the input to the robot, (as occurs in exptB), could cause any putative phenomenal states to differ from those it experienced in exptA; with the same input and initial conditions pertaining in each experiment the robot must execute the same computational state transitions and hence realise identical 'phenomenal states'.

However with a-priori knowledge of input to the robot we can collapse the contingent branching state structure onto an un-branching series of modal state transitions and hence perform the Putnam mapping onto any open physical system.

OBJECTION 2: FLETCHER, "THIS IS NOT CORRECTLY IMPLEMENTING THE FSA"

Putnam's mapping merely realises one specific series of state transitions, a particular FSA execution trace, and does not capture the full power of the FSA. To illustrate this consider the following experiment.

Consider a FSA to recognise a string in a given language. Just getting answer right once is not enough to say of the FSA that it recognises the string. What matters is the sequence of states the machine would enter if it had been presented with other strings[4].

But this conflates the clearly functional property, 'recognition of a string', with the clearly experiential property of 'instantiating genuine phenomenal states'.

If the property of 'recognition of a string' is only positively defined for an FSA that gets it right in all contingencies, then clearly it will be necessary to implement its full contingent FSA structure. However, no such definition is implicit in the notion of experiencing a phenomenal state[5].

OBJECTION 3: CHALMERS, "LACK OF COUNTERFACTUALS"

It is obvious that Putnam's mapping does not reproduce a full isomorph of an FSA with input; in particular it lacks ability to correctly implement counterfactuals. This lack of input sensitivity/counterfactual-behaviour is extremely significant hence there is no sound reason to suppose that the phenomenal states of two robots - one controlled by an open physical system and a suitable Putnam mapping; the second controlled by a FSA executing a suitable 'Artificial Consciousness program' – would be the same; if so, the DWP reductio must fail.

However, consider two experiments in which just such two robots are asked to report the colour of say, a bright red square presented as input:

- [Racp] is controlled by a FSA executing a putative 'Artificial Consciousness Program'.
- [Rput] is controlled by an 'open physical system' and suitable Putnam mapping.
- Input is identical (Xacp = Xput = 'a bright red square').

Now imagine building a large number of robots, [Racp .. R_n .. Rput], which serve to morph [Racp] into [Rput] by incrementally replacing each branching state transition in Racp, with a linear state transition, (contingent on the current input), in Rput, e.g.

IF (I > 0) THEN {A} ⇒ {B} ELSE {A} ⇒ {C}

Given input (I = 1) the above contingent state transition simply reduces to {A} ⇒ {B}.

COUNTERFACTUALS CAN'T COUNT

Now consider the putative phenomenal experience of R_n - what is it like to be R_n? If Rput does not have phenomenal experience as Chalmers claims then R_n's experience must either gradually fade, (e.g. say from bright red, to tepid pink to nothing), or suddenly disappear at some point.

But either case implies the mere removal of a section of the FSA state structure that, given the known input, is not and never could be entered, somehow influences the phenomenal states experienced by the robot. And conversely the mere addition of a segment of [nonsense] FSA structure that, given the known input, is not and never could be entered, would equally affect the robot's phenomenal experience...

Hence the phenomenal states experienced by [Racp] and [Rput] must be the same; ***counterfactuals cannot count***.

MECHANICAL BODIES – MYTHICAL MINDS?

If [Racp] experiences phenomenal states as its program executes then so must [Rput]. But if [R_{PUT}] experiences phenomenal states then Panpsychism is true, because, using the Putnam mapping, we can generate the appropriate modal state transitions in any open physical system.

Thus, via the reductio, [Racp] cannot experience genuine phenomenal states purely in virtue of executing a particular series of modal state transitions and the Artificial Consciousness project must fail.

CONCLUSIONS

In his 1992 book, 'The Rediscovery of Mind' Searle suggests that, "The study of the mind is the study of consciousness, in much the same sense that biology is the study of life" and concludes that "consciousness is a prerequisite for mental states", since via the Connection Principle: "… any mental state must be, at least in principle, capable of being brought to conscious awareness", (ibid).

Hence, since the DWP reductio suggests that genuine phenomenal states are not instantiated by the mere execution of any computer program, 'machines' are therefore incapable of carrying genuine mental states purely in virtue of executing the appropriate program and any computational account of mind must ultimately be found lacking.

So although it is time, as Chalmers suggests, (1996b), to "take Consciousness seriously", the mystery of consciousness is not explained by the execution of any computer program, for the DWP reductio demonstrates that if a computer instantiates consciousness purely in virtue of executing a program, then

[4] Personal communication, Dr. Peter Fletcher, Dept. Computing, University of Keele.

[5] Even if it were the case that to feel pleasure it is necessary to feel pain; there is no reason to suppose that one, who for some reason could not feel pleasure, when given a painful stimulus, would not still experience something.

consciousness is all pervading and little pixies are dancing everywhere.

REFERENCES

Bishop, J.M., (2002a), *Dancing with Pixies*, in Preston, J. & Bishop, J.M., (eds), *Views into the Chinese Room*, (Oxford: Oxford University Press).

Bishop, J.M., (2002b), *Counterfactuals Cannot Count: a rejoinder to David Chalmers*, Consciousness & Cognition, **11(4)**, pp: 642-652.

Chalmers, D.J. (1994), *On Implementing a Computation*, Minds and Machines, **4**, pp.391-402.

Chalmers, D.J. (1996a) *Does a Rock Implement Every Finite-State Automaton?*, Synthese, **108**, pp.309-333.

Chalmers, D.J. (1996b) *The Conscious Mind: In Search of a Fundamental Theory*, (Oxford: Oxford University Press).

Harvey, I, (2002), *Evolving Robot Consciousness: The Easy Problems and the Rest*, in Fetzer, J.H., (ed.), *Evolving Consciousness: Advances in Consciousness Research Series*, (Amsterdam: John Benjamins).

Hofstadter, D.R., (1981), *Reflections on Minds, Brains & Programs*, in Hofstadter & Dennett, (eds), *The Mind's I*, (Penguin).

Kurzweil, R., (1998), *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, (New York: Viking).

Minsky, M., (1985), *The Society of Mind*, (New York: Simon & Schuster).

Moravec, H.P., (1988), *Mind Children: The Future of Robot and Human Intelligence*, (Cambridge, MA: Harvard University Press).

O'Regan, J.K. & Noë, A., (2001a), *A sensorimotor account of vision and visual consciousness*, by, Behavioural and Brain Sciences, **24(5)**, pp. 939-1011.

O'Regan, J.K. & Noë, A., (2001b), *Authors' Response: Acting out our sensory experience*, Behavioural and Brain Sciences, **24(5)**, pp. 1011-1031.

Putnam, H., (1988), *Representation & Reality*, (Cambridge MA: Bradford Books).

Penrose, R., (1989), *The Emperor's New Mind: concerning computers, minds and the laws of physics*, (Oxford: Oxford University Press).

Penrose, R., (2002), *Consciousness, Computation*, in Preston, J. & Bishop, J.M., (eds), *Views into the Chinese Room*, (Oxford: Oxford University Press).

Searle, J.R., (1980), *Minds, Brains & Programs*, Behavioral & Brain Sciences, **3**, pp.417-424.

Searle, J.R., (1992), *The Rediscovery of Mind*, (Cambridge, MA: MIT Press).

Warwick, K., (1996), *Prospects for machine consciousness*, RSA, (Royal Society of Arts), Journal, June 1996, pp. 47-51.

Warwick, K., (2002), *Alien Encounters*, in Preston, J. & Bishop, J.M., (eds), *Views into the Chinese Room*, (Oxford: Oxford University Press).

Weizenbaum, J., (1976), *Computer Power and Human Reason: From Judgement to Calculation*, (San Francisco: W.H.Freeman).