Mechanical Bodies; Mythical Minds: dancing with pixies?

"*the action of a machine … seems to be there in it from the start*",
Ludwig Wittgenstein, Philosophical Investigations, (193).

Dr. J.M.Bishop
- Reader in Computing
Goldsmiths College,
University of London.

The Artificial Consciousness debate: do androids dream of electric sheep?

- This talk does not address the question of whether a conscious machine will ever be built …
  - [Searle], 'Such machines already exist; we are such machines'
- … It simply addresses the question of whether a robot controlled by a suitably programmed computer, can ever be genuinely conscious **in virtue of executing an appropriate computer program?**
  - i.e. Can a [computationally controlled] robot genuinely experience [first person] sensations; phenomenal states of mind?
    - … pains, anger, sadness, the ineffable red of a rose etc.

Route map

- Two views of the Artificial Consciousness (AC) debate**.**
- A brief presentation of the '**Dancing with Pixies**' reductio…
- … leading to the following (modest ☺) conclusions:
  - **computation is not sufficient for consciousness experience or mentality (genuine mental states),**
  - **computationalism does not provide an adequate explanation of mind,**
  - **and the [computational] Artificial Consciousness project must fail …**

Torrance: "the barrier of consciousness ..."

- Many people hold the view that:
  - *"there is a crucial barrier between computer models of minds and real minds:* ***the barrier of consciousness****."* , (Steve Torrance).
  - 'Information-processing' and 'phenomenal (conscious) experiences' **are conceptually distinct**.
  - And there appear to be real problems with granting phenomenal (conscious) experience to computational systems **...**

Searle: consciousness - a prerequisite for mental states?

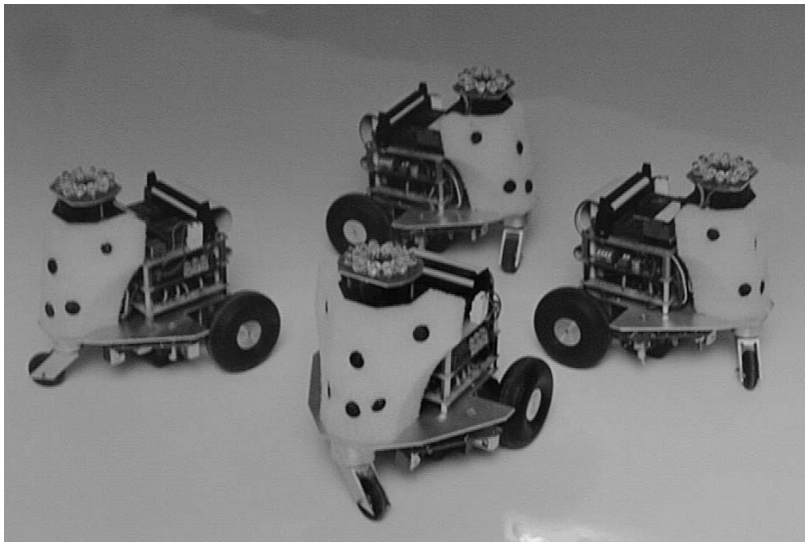
- "The study of the mind is the study of consciousness, in much the same sense that biology is the study of life", (Searle, *The Rediscovery of the Mind,* p 227).
- The Connection Principle: "… any mental state must be, at least in principle, *capable of being brought to conscious awareness",* (ibid).
- Hence, if machines are not capable of enjoying consciousness, they will be incapable of carrying genuine mental states …
- **… And computationalism must fail as a theory of mind.**

An opposing view …

- The continuity thesis:
  - machines can display genuine mental properties because …
  - **… Artificial Intelligence (AI) and Artificial Consciousness (AC) form an essential continuity.**
- In principle creating AC is no more problematic than creating AI.
- Systems displaying AC may already exist:
  - Because AC systems possess no properties which are not found in standard AI systems …
  - … they are just richer, more architecturally complex etc.

Kevin Warwick and the 'Seven Dwarves'

- The 1st generation
  - Simple reactive mobile robots.
- The 2nd generation
  - Cybernetic 'learning' robots
  - "As conscious as a slug"
    - *Prof. Kevin Warwick*



Dancing with Pixies (DWP)

- DWP is a reductio ad absurdum that endeavours to demonstrate that:

1. Computational states are always relative to an observed function, i.e. they are not intrinsic to physical states of matter.

2. IF the following [assumed claim] is true, **"that an appropriately programmed computer instantiates genuine conscious states"**
3. THEN "**panpsychism holds**"!

4. *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 …*

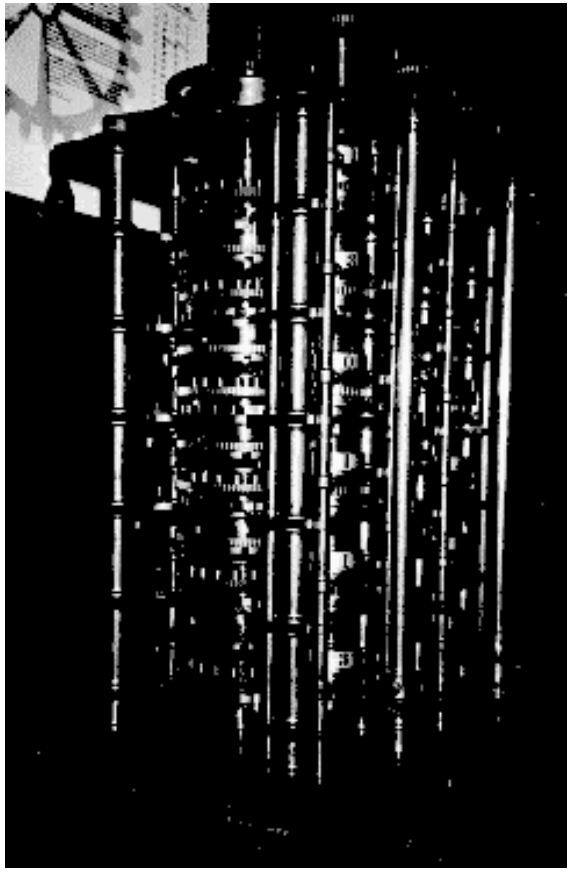
   … And we are led to reject the assumed claim, (2).

The core argument

- Computation is not an intrinsic physical property of matter.
- The computational processes of a 'conscious' computational system are characterised by a series of modal state transitions.
- The behaviour of any open system can be described by a series of modal state transitions.
- Over a finite bounded interval there exists a simple reliable mapping between these two domains.
- **Hence if a computer genuinely experiences phenomenal states then so does any open system. Panpsychism is true and disembodied minds, (little pixies), are to be found dancing everywhere …**

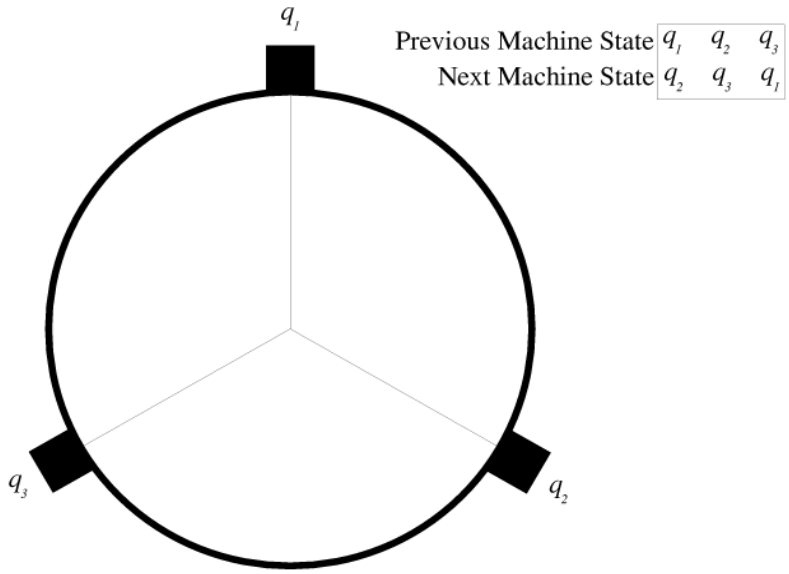
Computation is not intrinsic

- Computational states are always mapped onto physical states.
  - E.g. 5V = Logic TRUE.
- Hence computers do not have to be electrical
  - Babbage's Difference & Analytical Engines.
  - Weizenbaum's toilet roll computer.



Computational states are not intrinsic to physical states

- *Turing's 3-state input-less Discrete State Machine, DSM, (1950).*
  - At each time step the machine occupies one of a finite number of possible physical positions, (3).
  - In operation it cycles through a finite number of computational states, {Q_1, Q_2, Q_3}.
  - The next computational state is modally predicated by the current state.
  - The output (light on) occurs when machine in computational state Q_1.



TURING'S DISCRETE STATE MACHINE

- Turing's DSM in operation…

How a counter implements a 3-state input-less DSM

- 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 DSM (Q) generates the finite linear series of state transitions labelled, {Q_1, Q_2, Q_3, Q_1, Q_2, Q_3}.

- To implement the input-less DSM (Q) by the counter:
  - Map DSM state Q_1 to the disjunction (C_1 v C_4).
  - Map DSM state Q_2 to the disjunction (C_2 v C_5).
  - … and DSM state Q_3 to (C_3 v C_6).

- **As in any computational system the mapping assigns a logical, computational state onto a physical state of the system.**
  - By adopting this mapping any simple digital counter will generate the required state transition sequence, {Q_1, Q_2, Q_3, Q_1, Q_2, Q_3}, of the specified DSM over the defined time interval.

To implement an input-less Finite State Automata (FSA)

- The DSM-Counter mapping procedure can be easily extended to implement an input-less FSA.
  - To ensure all possible FSA states are generated, if the counter requires n states to generate the initial FSA state sequence then, if there remain un-entered FSA states, we increment the counter to counter state C_{n+1} and apply the mapping procedure again:
- An input-less FSA transits the sequences: $Q_1, Q_2, Q_3, Q_1, Q_2, Q_3$ AND $Q_4, Q_5, Q_6, Q_4, Q_5, Q_6$
  - Map $\{Q_1, Q_2, Q_3, Q\}$ as before.
  - Then map FSA state Q_4 to the disjunction (C_{n+1} v C_{n+4}).
  - Map FSA state Q_5 to the disjunction (C_{n+2} v C_{n+5}).
  - … and FSA state Q_6 to (C_{n+3} v C_{n+6}).
- NB. In general the above procedure is repeated until all FSA states are generated.

The physical **state** of things…

- In physics we can describe the time evolution of a complex system via a set of dynamic equations.

- By selecting appropriate intervals the system's behaviour can be quantised into a series of modal state transitions.
  - Hence a physical system can be characterised by a series of discrete states that evolve over time.
    - Eg. A four state quantisation of heating water in a kettle may be that the water goes from a cold state to a warm state; to a hot state; to a boiling state, (and perhaps eventually to a *cup-of-tea* state ☺).

- Any ***Open Physical System***, (e.g. a cup of tea), is characterised by a series of **non repeating** states that evolve over time, $\{S_1, S_2, S_3, S_4, S_5, S_6 \; .. \; S_\infty\}$.
  - Due to influence of cosmic rays, gravitational fields etc.
  - Analogous to an infinite counter, (i.e. one that never repeats itself).

The Tea God

- Consider that a cup of tea over the period [T_1 to 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 Super-mind, (e.g. 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 predict the future state of the cup of tea at any time.

- **Analogous to predicting the final state of a counter, given its initial state and the elapsed time, (e.g. number of clock ticks received).**

Putnam: 'Any open physical system implements any input-less 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. our 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).
  - **[Non-entered state sequences are generated as before].**

- **Once again, as in any computational system the mapping simply assigns a logical, computational state, onto a physical state of the system.**
  - By using this mapping any open physical system, (eg. a cup of tea), will also generate the required state transition sequence, {Q_1, Q_2, Q_3, Q_1, Q_2, Q_3}, over the specified time interval.

David Chalmers: on input-less FSAs …

- **Chalmers:** "… *the state-space of an input-less 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*"
  - As is demonstrated by the ease with which it can be implemented by an open physical system.

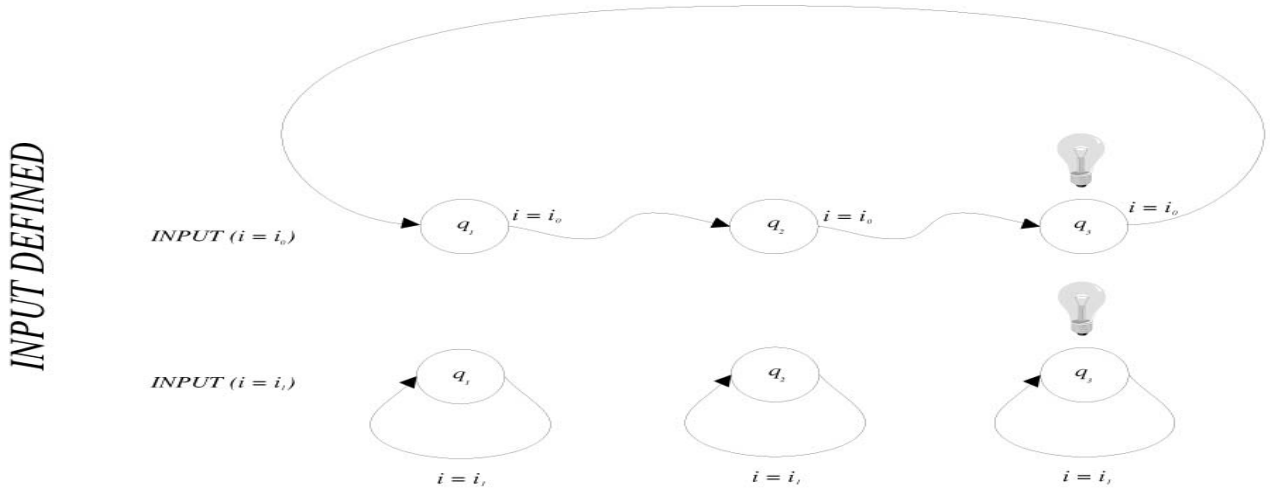
Automata with input and output

- To extend Putnam's result to deal with FSAs with I/O we need to give the open physical system memory to record all input to the system.
  - E.g. Make marks, or indentations, on the tea cup..
- Thus to implement the FSA :- for every possible input we map physical system state S_1^{abc} to the FSA state we get by starting at state Q_1 with input {abc}.
- Similarly, map physical state S_2^{abc} to FSA state Q_2 with input {abc} etc. etc. until all FSA states have been transit for all possible input strings.
- Thus, to implement a FSA with input, an exponential number of extra states are required by the open physical system:
  - If the input symbols are selected from an alphabet of size *k* and the length of the input string is *l* the extra number of counter/physical system states required compared to an input FSA is k^l.

Combinatorial explosion

- If we a assume a putative cognitive robot samples its environment for 8-bit input data at 100Hz …
  - … the number of states required to implement the FSA by the open physical system will grow at 256^{100} per second and rapidly become larger than the number of atoms in the known universe …
  - **… and the Artificial Consciousness project is preserved!**

FSA with fixed execution trace



- Knowledge of system input will collapse the contingent branching state structure of the FSA state transition diagram into a simple linear path.
  - Replace every contingent branch with a fixed modal transition defined by current state and input.
  - Hence, with input fixed, the state transition diagram becomes a simple un-branching sequence, analogous to that of an input-less FSA.

- Further, over any *finite interval*, all circular (iterative) paths can be unfolded to produce a finite linear series of state transitions.
  - *Lock off in Q_1 {Q_1 Q_2 Q_3}; Lock on in Q_1 {Q_1 Q_1 Q_1}*

- **Hence, with its input fixed over a finite time period a '*Finite State Automata*' functions like a simple clockwork device …**

Happy machines?

- So can a machine, (robot), experience phenomenal states in virtue of executing a program?
  - Without loss of generality define the robots behaviour by FSA $Q_{(p, x)}$
    - An FSA (Q) executing its state transition table, *or program*, (p) on **defined (**fixed) input (x).
  - Consider the action of $Q_{(p, x)}$ over the interval $[T_1 .. T_n]$
    - As input fixed state transition diagram unfolds to a linear path.
    - I.e. $Q_{(p, x)}$ generates a finite linear series of state transitions at clock intervals of Q, $\{Q_1, Q_2, \ldots Q_n\}$.
  - It is the claim of **Artificial Consciousness Researchers** that during this time interval, as the robot executes its program, **p**, genuine phenomenal states, (e.g. happiness), are 'mechanically' realised.

Happy tea?

- Is a cup of tea happy?
  - Over the observed time period $[T_1 .. T_n]$ map 'cup of tea' states $\{S_1, S_2, .. S_n\}$ to robot states $\{Q_1, Q_2, .. Q_n\}$, using the Putnam transform.
  - Just as for the robot, the 'cup of tea' state transitions are modal:
    - Given initial conditions $\{S_1\}$ and boundary conditions, we can predict any future state $\{S_n\}$.
  - 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 a particular 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"

- This is not a real mapping as we can only perform Putnam style mappings a-posteriori once we know the input(s) to the robot.
  - " ***This this is not science!*** "
  - But we can simply repeat the experiment using exactly the same input to the robot …
    - The robot is fully deterministic; by what **scientific procedure** can simple a-priori knowledge of its input, impact the phenomenal states it purports to experience over the specified time interval?
  - **… yet with a-priori knowledge of input to the robot we can effectively collapse the contingent program state structure into an unbranching series of state transitions and hence perform the Putnam mapping onto any open physical system; thus the DWP reductio holds.**

Objection 2: Fletcher, "This is not correctly implementing the FSA"

- Putnam's mapping merely realises a desired series of state transitions and does not capture the full power of the FSA.
  - Fletcher, "Consider a FSA to recognise a string in a given language"
    - Just getting answer right once is not enough to say that the system recognises the string.
    - What matters is the sequence of states the machine would enter if it had been presented with other strings.
  - But this conflates 'recognition of a string' with 'experiencing phenomenal states'.
    - It may be the case that to say of the FSA that it correctly 'recognises a string' it is necessary to implement its full structure; but Fletcher's objection does not prove that such 'full structural implementation' is necessary for the system to realise phenomenal experience ...

Objection 3: Chalmers, "Lack of Counterfactuals"

- "Fixing input implies the system is not isomorphic to a FSA; in particular it lacks ability to correctly implement counterfactuals."
  - Consider two robots being asked to report the colour of a bright red square.
    - [R_{SAI}] is controlled by an 'Artificial Consciousness program'.
    - [R_{PUT}] is controlled by any Putnam style, 'open physical system'.
  - Now imagine building a series of robots, [R_{SAI} .. R_n .. R_{PUT}], morphing [R_{SAI}] into [R_{PUT}]
    - Incrementally replace each branching state transition in R_{SAI}, with a linear state transition, (contingent on the current input), in R_{PUT}.
    - IF (I>0) THEN {**STATE-A** / e.g. lamp on} ELSE {**STATE-B** / lamp off}
      - E.g. Given input (**I=1**) this counterfactual becomes {**STATE-A** / lamp on}
- **Now, what is it like to be R_n?**

But 'counterfactuals can't count'

- If R_n does not have phenomenal experience then either it must gradually fade, (eg. bright red .. tepid pink .. nothing), or suddenly disappear at some point.

- But either case implies that the mere deletion of a sequence of possible state transitions that, **given the input cannot and is not executed**, will somehow influence the phenomenal states experienced by the robot …

  - … which would also conversely imply that, the mere addition of a sequence of [nonsense] state transitions, that could never be entered, would also affect the robot's phenomenal experience!

- **Hence the phenomenal states experienced by [R_{SAI}] and [R_{PUT}] must be the same; counterfactuals cannot count …**

The artificial consciousness project is dead?

- If [R_{SAI}] experiences phenomenal states as its program executes then so must [R_{PUT}] …
  - … and if [R_{PUT}] experience 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, [R_{SAI}] cannot experience genuine phenomenal states purely in virtue of executing a particular series of modal state transitions and the Artificial Consciousness project must fail.**

Dancing with Pixies …

- Although it is certainly time to, "**take Consciousness seriously**", the mystery of consciousness is not explained by the execution of **any** computer program …

  - … ***for if a computer instantiates consciousness purely in virtue of executing its program, then consciousness is all pervading and little pixies are dancing everywhere***.

- **For further details see:**
  - Bishop, J.M., (2002), *Dancing with Pixies*, in, Preston, J. & Bishop, J.M., (eds), *Views into the Chinese Room*, OUP, Oxford, UK.

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