How many thoughts can you think?

Richard Rohwer
Dept. of Computer Science and Applied Mathematics
Aston University, Birmingham B4 7ET, UK

20 November 1992

Abstract

In ordinary computer programmes, the relationship between data in a machine and the concepts it represents is defined arbitrarily by the programmer. It is argued here that the Strong AI hypothesis suggests that no such arbitrariness is possible in the relationship between brain states and mental experiences, and that this may place surprising limitations on the possible variety of mental experiences.

Possible psychology experiments are sketched which aim to falsify the Strong AI hypothesis by indicating that these limits can be exceeded. It is concluded that although such experiments might be valuable, they are unlikely to succeed in this aim.

1 The argument in brief

The Strong AI hypothesis [2], is that a mind is associated with any suitably sophisticated machine [1]. Given this, it seems reasonable to presume furthermore that any specific mental experience is associated with (or 'produced by') one or more specific machine states or state trajectories. The number of distinct possible mental experiences must therefore be limited by the number of possible machine state trajectories.

The set of possible machine state trajectories may be a continuum, and therefore infinite, but it can be reasonably argued that a finite subset of these is entirely representative so far

[1] Strong AI asserts that algorithms underlie consciousness. Here a less specific hypothesis is adequate – that consciousness arises from certain patterns formed in the brain's activity.

as cognitive function is concerned. Furthermore, one can easily argue that this subset must be vast compared to any plausible estimate of the number of distinguishable thoughts and experiences contained in a human lifespan. But if the relationship between state trajectories and the associated mental experiences is fixed by nature, then this finiteness applies to the *thinkable*, and not merely to the *thought*. That is, if at least one brain state trajectory has to be allocated to every possible thought or experience, regardless of who, if anyone, ever actually has that thought or experience, then everyone shares just one finite repertoire of possible thoughts and experiences with which to decorate their life.

A finiteness constraint on the thinkable is somewhat disturbing. It is possible to invent schemata based on power sets of power sets, etc., which contain arbitrarily vast numbers of elements, numbers like $2^{2^{2^{2^{2^2}}}}$. Of course no one has time to think about every element of such a schema individually, but finiteness of the thinkable implies that it is impossible to think of some elements of the schema at all. This is so because given the (finite) number of possible thoughts, it is an easy matter to invent a schema which contains more elements than that. Thus, there is an element which is impossible to think of, even if one is free to direct full attention to it, and every possible effort is made to point it out. Put another way, there are pairs of elements which cannot be consciously distinguished, no matter how much effort is put into recognising the distinction.

Perhaps this conclusion is acceptable. Perhaps the thinkable is a finite set, and psychology experiments can provide estimates of its size by measuring the human ability to distinguish elements of large schemata. The estimates of the size of the cognitive state space might even be combined with neurophysiological data to estimate the length and time scales of cognitively significant patterns of brain activity.

Perhaps this conclusion is not acceptable. One way out is to insist that there is a sense in which infinitesimally distinct brain states have distinct mental significance, thus allowing the thinkable to be infinite. Or perhaps Strong AI is wrong; brain states are not the sole determiners of mental states.

2 The argument in detail

In this section, the argument for finiteness of the thinkable is expressed more formally, and rough estimates are given for the relationship between the scale of cognitively relevant components of brain-state patterns and the size of the thinkable.

2.1 Notation and Terminology

The notation and vocabulary used in [1] is convenient for the present purposes. That paper formalises the representation of arbitrary concepts by states of neural network models, and the same formalism serves to represent the support of mental states by brain states.

Let us assume that a brain and its functioning can be fully described by a time-dependent array of numbers $\mathbf{y}(t)$. This is difficult to dispute without affront to some of the soundest theories known, which hold that any physical system can be fully described by its quantum state. This is just a complex vector, usually of infinite dimension. But if the 'mind-magic' arises from the computational aspects of the brain embodied in the firing patterns of the neurons, then appeal to the most fundamental physical variables is an unwarranted distraction. A derived set of variables minimally describing the firing pattern at any time would serve better for $\mathbf{y}(t)$. In any case, let us call a segment of a state trajectory over a finite time T, $\{\mathbf{y}(\tau)\}_{\tau=t-T}^{\tau=t}$, a *state excursion* or simply an *excursion*. Let us call any mental phenomenon arising from an excursion an *experience*, and say that the excursion *represents* the experience.

Of course, unlike computers, brains are not produced on a tightly-controlled assembly line, so the hardware differs somewhat from one brain to the next. This provides an argument against using just one vector space to represent all brain states. So to proceed it is necessary to assume that a vector space big enough to represent an arbitrary brain is not much larger than a space big enough to represent just one; *ie.*, that there is much commonality between brains in their cognitively-significant structure, so comparatively few variables are needed to express the differences. In the extreme, this could be defeated by insisting that variables representing the position of each brain component in space are cognitively important, so that structurally identical brains in different locations could have different cognitive possibilities.

This is a somewhat odd use of the term 'represent'. It would be quite normal to say that some of the patterns of electrical activity in the components of a conventional computer 'represent' certain concepts which a programmer has arbitrarily specified. The Strong AI hypothesis (as taken here) asserts that the 'representation' of an experience by an excursion is not an arbitrary relationship, but entails something along the lines of cause and effect. Therefore it might sound better to speak of experiences 'caused by' or 'supported by' excursions of brain states, and concepts 'represented by' excursions of computer states. But this would emphasise the distinction which Strong AI seeks to blur. The objective of Strong AI is to put these types of statements on a more equal footing, so a uniform vocabulary and notation seems worth the oddities it may bring with it.

For any experience A let there be a like-named Boolean-valued *test map* which can be applied

to any excursion Υ with the significance:

$$\begin{aligned} A(\Upsilon) &= 1 \quad \text{means '}\Upsilon\text{ represents }A\text{'} \\ A(\Upsilon) &= 0 \quad \text{means nothing.} \end{aligned} \tag{1}$$

Assigning a null interpretation to the 0 case makes it possible to have test maps which avoid borderline cases. The price for this convenience is that a separate test map $\overline{A}$ must be introduced to positively assert non-representation of A; one cannot automatically assume that $\overline{A}(\Upsilon) = 1 - A(\Upsilon)$. The set of excursions which represent experience A includes (but might not equal) $A^{-1}(1)$, and the set of experiences represented by Υ includes (but might not equal) $\{A|A(\Upsilon) = 1\}$.

If it is desired to support the notion that an experience may be represented more strongly by some excursions than others, then a different test map can be introduced for every strength or quality of representation envisaged.

2.2 Size of excursion set

Although the number of neurons and synapses are finite in number, the representationally relevant variables might still include continuous ones such as firing strengths and the relative spatial positions of the various brain components. It seems a rather desperate, however, to load each of the infinite bit strings afforded by continuous variables with distinct representational significance. It would be remarkable if arbitrarily small changes of such variables should result in perceptible changes in the mind, not least because the mind's robustness would be difficult to explain. Therefore let us make a set of assumptions which effectively forces these variables to take values from a finite set.

An obvious way to proceed is by cutting up a finite-volume state space into cells based on finite values ϵ_{state} and ϵ_{time} for cognitively meaningless variations in state variables and time. Let the state space have finite dimension N, and let the components of the state vector be real numbers scaled to lie between 0 and 1, $0 \leq y_i \leq 1$. Similarly, let the unit of time be the longest time that needs to be used to represent a distinct thought. Then the number of distinct cognitively meaningful states is something like $\left(\frac{1}{\epsilon_{\text{state}}}\right)^N$ and the number of excursions is $\left(\frac{1}{\epsilon_{\text{state}}}\right)^{NT}$.

2.3 Remarks on the form of the separation assumptions

It is necessary to be somewhat careful with the phraseology of the assumption which justifies introducing the scales ϵ_{state} and ϵ_{time}, in order to disallow a change of cognitive state as a brain state is perturbed infinitesimally at a cell boundary. This section gives one way of going about it.

Let us say for excursion Υ, *attention is focused* on experience A *more than* experience B if $A(\Upsilon) = 1$ and $\overline{B}(\Upsilon) = 1$. It is possible to focus attention on A more than B if

$$A^{-1}(1) \bigcap \overline{B}^{-1}(1) \neq \emptyset. \tag{2}$$

This denotes a set of excursions which represent A and not B.

Let $\Upsilon_A = \{\mathbf{y}_A(\tau)\}_{\tau=t-T}^{\tau=t}$ be an excursion which represents attention focused on A more than on B: $\Upsilon_A \in A^{-1}(1) \bigcap \overline{B}^{-1}(1)$. Select Υ_B conversely: $\Upsilon_B \in B^{-1}(1) \bigcap \overline{A}^{-1}(1)$. Let us assume that any such Υ_A and Υ_B are not related by any finite state-time transformation limited by *state warp* constant ϵ_{state} and *time warp* constant ϵ_{time}. By this it is meant that for any function $\chi(\tau)$ with range $[-1, 1]$,

$$|y_{Ai}(\tau) - y_{Bi}(\tau - \chi(\tau)\epsilon_{\text{time}})| > \epsilon_{\text{state}} \tag{3}$$

for each component i of $\mathbf{y}_A$ and every time $\tau \in [t - T, t]$.

This has the desired effect of keeping representations of distinct experiences finitely separated without setting up precisely located cell boundaries.

It might be tempting to try to achieve this by making the simpler assumption that for any Υ_B within a small state warp and time warp of Υ_A, that $A(\Upsilon_B) = 1$; *ie.*, if Υ_A represents A then any nearby excursion also represents A. But this would be disastrous, because it would be possible to conclude that every excursion represents every experience, by considering a succession of small transformations along a continuous path between a pair of excursions representing an arbitrary pair of experiences.

3 Numerical Estimates

Let us estimate the number of excursions which can represent distinct mental experiences. The dimension of $\mathbf{y}$ can be estimated as $N = N_{\text{cells}} N_{\text{cellparts}}$, where N_{cells} is the number of

neurons in the brain and $N_{\text{cellparts}}$ is a typical number of representationally-significant components of each neuron. The latter might be the number of synapses per neuron (say 10^4) or the number of finite-elements needed in a detailed electrical model of the neuron (say 10^9). Or it might be as small as 1, if it is only whether the neuron is firing that matters. The number of neurons may be something like 10^{12}, but perhaps only 10^9 of them are directly involved in representing experiences. Let N_{levels} be the number of representationally-significant levels of electrical activity which can occur in each cell component, perhaps 100; perhaps 2.

Let us regard any experience to be confined to a psychological moment of time. Even if the experience involves distant memories, it is the representation of the recollection of those memories at the particular moment that counts, not the representation of the original experience being recalled. A psychological moment of time may be spread over as much as a second, and the fastest events likely to have representational significance might consume 1 millisecond. So the number of time steps T in an excursion which represents an experience is at most 10^3, and at least 1.

The number of representationally-significant excursions is therefore

$$N_{\text{excur}} = N_{\text{levels}}^{N_{\text{cells}} N_{\text{cellparts}} T} \tag{4}$$

which might be as much as $100^{10^{12+9+3}} \approx 10^{10^{24}}$ or as little as $2^{10^{9+0+0}} \approx 10^{10^9}$.

The maximum might be increased still further by attaching representational significance to configurational degrees of freedom in the brain, in addition to the degrees of freedom concerning activation. For example, there are roughly $2^{N_{\text{cells}}^2} = 2^{10^{24}}$ ways for the neurons to be interconnected (neglecting any shortage of synapses for the purpose). But in any case there does not appear to be any way to compete with the arbitrary power sets whose elements might be thought to populate the thinkable.

4 Proposed Psychology Experiments

Let us examine more closely a schema which may include more elements than there are possible thoughts. Perhaps the most straightforward is the natural number system, the set of numbers between 1 and N_{excur} itself. These are at least billion-digit numbers, according to the above reasoning. So it would be interesting to know whether humans can distinguish any randomly selected string of a billion digits from any other.

One possible way to organise such an experiment would use 2-dimensional arrays of pixels, or a 3-dimensional array employing a colour dimension. Current commercial systems provide

on the order of 1000×1000 pixels with on the order of 1000 colours, just enough for a billion-digit experiment. Subjects would be shown a randomly generated pattern, and allowed to study it for as long as desired. They would then be shown, with 50% probability, either the same image again, or another which differs by 1 bit, and asked whether the new image were any different.

Strong AI looks relatively safe from attack by this sort of experiment, because it seems unlikely that humans will perform well in the billion-pixel case, and it would take strong performance on an awesome 10^{24} pixels to put the hypothesis into serious danger of disproof. Furthermore, this type of experiment only provides lower bounds on the size of the brain-state space for the part of the brain concerned with representing visual images, which in turn is only a lower bound on the number of thoughts or experiences the mind can have. Experiments to provide a measure of the entire variety of mental experience remain to be devised.

Regardless of any implications for Strong AI, this type of experiment may provide evidence that representation of mental experiences from restricted domains is accomplished with neural machinery operating on particular spatial or temporal scales. If neurophysiological evidence suggests that a specific area of the brain is responsible for representing the experiences in question, then (4) can be applied to that area.

5 Conclusions

In a computer, the contents of memory have no *a priori* representational significance. A programmer has the freedom to impose an arbitrary interpretation on them. This freedom of interpretation allows a computer to represent many more ideas than the size of the state space of its memory. There is no analogous freedom in the relationship between mind and brain, if, as philosophies such as Strong AI suggest, mental experiences are underpinned by spacio-temporal patterns of neural activity in a specific, presently unknown, way. The size of the brain's state space places an upper bound on the number of possible mental experiences. This makes it interesting to examine large schemata of possible (or seemingly possible) experiences, such as those based on number systems or progressions of power sets, in search of a set of thinkable thoughts larger than the the set of brain states. Success in such an endevour would provide serious evidence against Strong AI.

Such an endeavour appears to be difficult, however. Psychology experiments seeking such a result may be interesting and valuable, but it seems unlikely that human performance in these experiments would be strong enough to threaten Strong AI.

References

[1] Richard Rohwer. A representation of representation applied to a discussion of variable binding. Technical report, Dept. of Computer Science and Applied Maths., Aston University, Birmingham B7 4ET, UK, 1992. To appear in Proc. Neurodynamics and Psychology Workshop, 22-24 April 1992, Bangor, Wales, M. Oaksford and G. Brown, Eds.

[2] J. Searle. Minds, brains, and programs. *Behavorial and Brain Sciences*, 3, 1980.