Philosophical Communications, Web Series, No 38
Dept. of Philosophy, Göteborg University, Sweden
ISSN 1652-0459

The essential connection between representation and learning

Helge Malmgren

Department of Philosophy, Göteborg University, Sweden

e-mail: helge.malmgren@filosofi.gu.se

http://www.phil.gu.se/helge/

Poster presentation at ASSC-10, Oxford, June 25, 2006

Four senses of “mental representation”

No doubt, we human beings often think about objects, remember facts and imagine events – in short, we are in different intentional states. In many of these cases we do the thinking, remembering and imagining without the help of any external representation. If nothing else is meant than the occurrence of such an un-aided intentional state by the phrase “mental representation”, it is quite uncontroversial that mental representations occur. To distinguish this sense from others, I will use the phrase *mental representing*.

Most would also agree that each occurrence of mental representing supervenes on an internal state of the human being which does the representing, possibly together with some fact about her/his environment. I will refer to such subvenient states as *mental-representing enabling states.* If nothing else is referred to by the phrase “mental representation” than an internal, mental-representing enabling state, it is again fairly uncontroversial that mental representations occur.

If, however, it is added that internal mental-representing enabling states are themselves mental in nature – i.e., that they are *mental representing-enabling states* – we begin to leave the arena of agreement. So, if the phrase “mental representations” is loaded with this idea, it is not obvious that such entities exist.

It is even less obvious and agreed upon that there are states, mental in nature and enablers of mental representing, which function like external representation in that the represented object is apprehended *via* an apprehension of such states. Mental representions in this sense, if they exist, I will call *indirect mental representations*.

Below, “mental representation” will only be used in the first and second of the above senses.

The simulation theory of mental representation

This recent update of Hume’s theory of ideas (another precursor is the so-called “stimulus-substitution” or S-S, theory of animal learning) says that the essential function of a mental represention is to work as a substitute for a perception when, for some reason or other, there is no perceiving. It is usually implied that any intentionality which is involved in a mental representation (other than a perception) is derived from the intentionality of the perception for which the representation substitutes.

No doubt, mental representing often fulfils this substitutive function well. Think of what happens when the light goes out on a winter night. We are then not left without all possibility to orient ourselves in the dark, but can use our memory images (so-called) to steer the course. Remarkably, such memory images often even share the *dynamic* properties of the perception which they substitute for. Just think of what happens when your radio is shut off before the end of a melody which you know by heart. The rest of the melody is reliably played in the intuition of your internal sense (using Kant’s terms).

If one asks for an explanation of these adaptive properties of many mental representations, an answer in terms of learning often lies near. I do not doubt that this kind of answer is often the best one. Neither do I doubt that evolutionary considerations are relevant to the issue what, in turn, explains the design of our learning mechanisms. But is learning an external add-on, so to say, to our representative abilities, and due only to evolutionary contingencies, or is there a more intimate connection? My answer is: The basic system design that makes it at all possible for mental representations to take the role of perception also explains why these representations so easily adapt to environmental constraints – i.e., why they are so often *successful* substitutes. The present poster is an attempt to spell out and argue for this answer.

A preamble: habituation as a natural property of polystable systems

Habituation is a widespread and important learning phenomenon and a discussion of it will function well as an introduction to the kind of abstract argument upon which the rest of this paper will build.

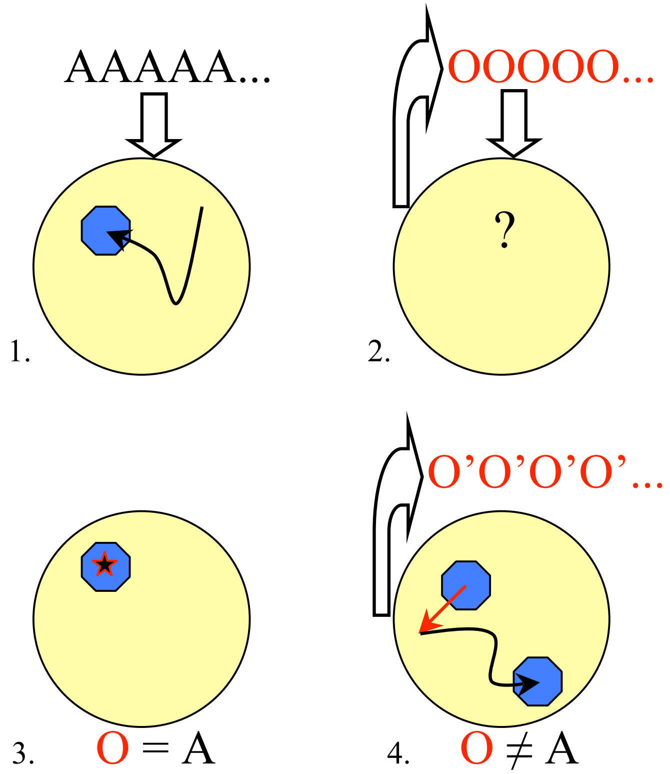
In his classic *Design for a Brain* (1952), W.R. Ashby argues that the phenomenon of habituation "is to be expected to some degree in all polystable systems when they are subjected to a repetitive stimulus or disturbance" (p. 189), and that this very general system property is the common factor of the different, detailed explanations for habituation in different kinds of organisms (actually, it is found in the amoeba as well as in man). By a "polystable" system Ashby refers to a deterministic machine whose parts have many equlibria and are, in a certain sense, pseudo-randomly joined. His argument is only sketched but in (Malmgren 1984), I made it more explicit and proved his point for a certain class of systems:

Take the transition table of a finite deterministic automaton with n states and m outputs, and fill it with uniformly random integers from 1 to n. The result will be a *randomly composed automaton.* Think of it as an ensemble of $(nm)^n$ different automata, and of its behaviour over time as the behaviour of this ensemble. Imagine an organism whose brain is actually a large sample of the randomly composed automaton, that all subsystems receive the same input, and that the change in some global output state of the organism is proportional to the sum of all state changes in the subsystems. Now let the organism receive a repetitive input. With a probability of 1/n, each subsystem will go to a point attractor under the first input, which means that it will not change state any more. Of those systems which do not go to a stable state at the first moment, a fraction of 1/n will do so at the next step, and so on. During the first n time steps, the amount of global change of the whole organism will therefore decline gradually.

Introducing representations by means of feedback

A point which is not always made in connection with the simulation theory of mental representation is that in order to substitute for perceptions in the organism's brain machinery, the representations must be made available as inputs at some level of that machinery. As the representations are internally produced, it follows that there must be a feedback loop. This might strike the reader as an extremely trivial point. What is not trivial, however, is that the existence of such a feedback loop in a polystable system *automatically* results in adaptive representations, and therefore of learning – including associative learning, and learning of sequences. This is what the rest of this poster is about.

The basic logic of the situation is actually not much more complex than in the case of habituation. Let us discuss it using Ashby's terminology. Suppose (1) that a polystable system with feedback is given a repetitive input A, alternating with periods of no input. During the latter periods (2), the output O of the system is instead made available as input. What will happen? Well, *if* this output/input O is equal to A, the system will for the rest of the day remain in any stable state which it had reached during input A (3). If O ≠ A, the system will (with a finite probability) go to another point in its state space (4). From there, it may well find a new attractor under input A, in which its output O' is actually A. If not... and so on. In short, the system will tend to learn to represent the input correctly.



Sequence learning and conditioning

It can also be shown that polystable systems – here represented by the randomly composed automaton – will also tend to form adaptive *sequential* and *conditioned* representations. We now suppose that the randomly composed system also has a randomly assigned output function. It is given a recurrent sequential input, for example ABCABC..., and after that its own output as input. Then there is a non-zero probability that the system remains stable in the latter circumstance because ABCABC... (in phase with the first input) is the output of the system when it is in a limit cycle to which it went under the external input. If not... the rest of the argument should now be obvious.

The case of conditioning is perhaps more interesting, especially when the conditioned stimulus is a sequence of inputs. Let us repeatedly present AB followed by C, this whole sequence being interlaced with long sequences of a background stimulus D, to the randomly composed machine. Then we test it with AB and the output of the machine as its next input. *If* the system has entered a stable behaviour cycle such that it always returns to the same state on the "B", an output after AB in the form of a C will keep it in this attractor for the rest of the day. Any other output will tend to displace the system to another region of the state space – again *etcetera.* In other words, the system will over time tend to give the conditioned response C to the sequence AB. Furthermore, this response will be specific to AB.

It might seem a surprising claim that classical conditioning has a simple system-theoretic explanation of this kind. An equally simple simulation (Malmgren 1991, 1997) demonstrates both the validity of the argument and the fact that the effect can be quite significant already in small systems, at least if one adds some general stability in the form of a low tendency for stimulus D to change the system's state.

Random walks and directed search “strategies”

The use of randomly composed finite systems as models of non-associative and associative learning entails several simplifications and limitations. Perhaps the most important *simplification* is that finite-state systems do not have strange (chaotic) attractors, but many continuous systems have. Hence great care has to be observed when generalising from finite to continuous systems using the attractor concept. But if one limits oneself to discussing non-chaotic systems, or non-chaotic regions of any system, the above arguments from the polystability of randomly composed finite systems surely have analogues in terms of continuous-valued systems. There is however not room to make these analogies explicit here.

One important *limitation* of the randomly composed finite system is the very randomness of its design. Almost the only thing you know about it is that it has point attractors and limit cycles, but nothing is known about the paths which the system takes to these attractors. In terms of learning, this means that the solutions are stable when found, but they are found in a completely random fashion. So, one could say that the randomly composed finite automaton performs like a dedicated Popperian.

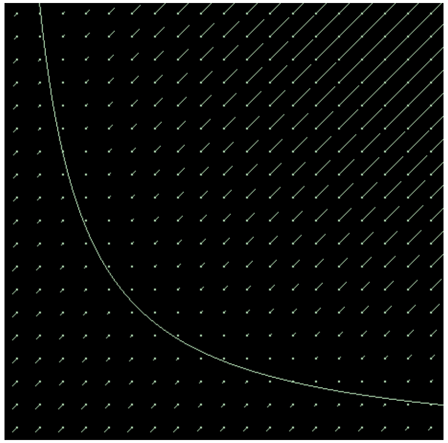
In an earlier poster (Malmgren 2002), it was shown that a certain class of continuous systems with feedback will indeed perform directed searches for the best solutions to learning problems. They will therefore rather behave as Baconian inductive reasoners. The reader is probably already familiar with one such class of systems, viz. (continuous-valued) neural networks with adjustable connections strengths and error-minimizing learning algorithms. However, the class characterized in the earlier work is (in an important respect) much more general and can for example account for learning on the cellular level. Instead of repeating the general results here, I will present two illustrative examples.

The learning thermostat

Suppose that you don't only want to regulate the temperature of a room so that it keeps close to a certain level, but want to regulate it to the same temperature as a certain *other* room. For this, I suggest using a thermostat which has the following extra function: if you place it in a room where the temperature is kept constant by other means, its internal setting reliably adapts itself to that temperature. (Also, that setting can be held constant by means of a certain switch.) – OK, how does one build such a device?

We have to do with a system with two variables, the external temperature E and the internal state I. To guarantee stability, we want the system to contain a (connected) set of temperature/setting pairs which are point attractors for both variables simultaneously. Moreover, for each such point $\{E_0, I_0\}$ the system should move smoothly towards that point if *either* E or I is kept constant at their respective value E_0 and I_0. One can summarize these requirements by saying that this whole set of points must form a *learning continuous attractor* (Malmgren 2002). Here is one solution:

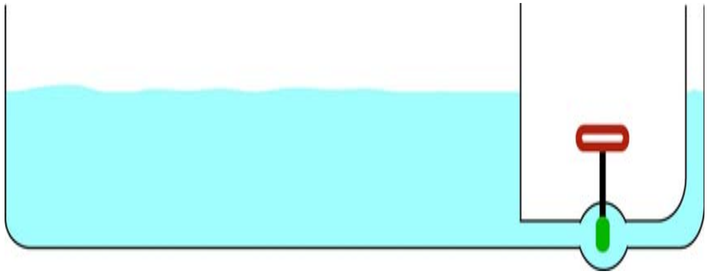
If $dE/dt = dI/dt = 1 - ET$, then the set of points satisfying the equation $I = 1/E$ will form a learning continuous attractor for the system. Keeping the external temperature (but not the internal state) constant at any value E_0 will make the thermostat gradually change its state to $I_0 = 1/ E_0$; letting it regulate a room using this state will, conversely, result in temperature E_0. (Letting the system completely loose will make it follow the flow depicted in the figure to the right.)



What has representation to do with similarity?

In terms of the simulation theory of mental representation, the output E produced by the learning thermostat when in the "regulating" mode should be regarded as the analogue of *representing,* while the internal state I is the analogue of the *representing-enabling state.* The former, *but not the latter,* is similar to "perception" (i.e. external temperature E when the device is in the learning mode). Indeed, similarity between internal state and input/output *can* be used to build a learning continuous attractor. This was exploited by Seung in his model of two identical, linear neurons with mutual excitation (Seung 1996). Here, the learning continuous attractor is the identity function E = I. Now, it can be shown (Malmgren 2002) that the identity function is the *only* possible choice for a monotonically *increasing*, symmetric learning continuous attractor, while there are innumerable *decreasing* suitable functions (most of which are non-linear like I = 1/E). My guess is that the brain often codes signals using different versions of such symmetric but "inverse" codes, just as we can code a photograph by its negative, or a casting by its mould. So, even if "pictures in the head" are rare, there may be lots of negatives in there!

Our second simple example of a learning continuous attractor uses the identity function. The level of water in the big compartment is the internal state while the input is the level in the narrow tube. The large basin can be filled to any desired level through the tube, and will later be able to reconstruct that level. A similar, slightly more complex model (Malmgren 2002) involving diffusion over membranes might explain cellular memory (Egorov et al 2002, Fransén et al 2006).



But perception is not an input!

The philosophically oriented reader may object to all the above arguments by pointing out that perception cannot really be looked upon as the *input* to the brain-system. Isn't perception rather the end product of the brain's processing of the input? And if so, what have the abilities of certain kinds of systems to duplicate their *input* to do with our ability to simulate *perception*?

No one can be more happy with this objection than the author. I readily agree that the above arguments have used a gross simplification of reality. However, they can all easily be changed into more correct analogies with simulation of perception, namely if perception is regarded as the *output* of the brain-system under certain ("perceiving") circumstances – an output which is *also* fed back to the system and thereby *becomes* an input. Now, what has to be shown is only that a polystable system's fed-back outputs under "perceiving" and "non-perceiving" circumstances tend to adapt to each other.

Indeed, not only is it obvious that similar arguments to those given above are valid for this reformulation of the problem. It now also becomes possible to frame Ashby-style explanations of the effects of (other) mental representing *on perceiving* – not only the other way round. Furthermore, it becomes much more comprehensible on an intuitive level how the representings, which we formerly looked upon as the "outputs" of the system, can even attempt to substitute for that which we regarded as its "inputs" (the perceivings). For how is the *output* coupled to the *input* channels? But if both are outputs that are fed back to the system, it would rather be a mystery if they did not use the same channel(s) for the feedback.

By the way, the most important of these feedback channels is better known as "consciousness".

Acknowledgements

I am very grateful to Björn Haglund for years of cooperation on basic problems of learning, and for comments on a preliminary version of this poster. Thanks also to Erik Fransén for stimulating discussions about cellular memory.

References

Ashby WR (1952), *Design for a Brain.* Chapman & Hall, London.

Egorov AV, Hamam BN, Fransén E, Hasselmo ME, Alonso AA (2002), Graded persistent activity in entorhinal cortex neurons. *Nature 420,* 173-8.

Fransén E, Tahvildari B, Egorov AV, Hasselmo ME, Alonso AA (2006), Mechanism of graded persistent cellular activity of entorhinal cortex layer V neurons. *Neuron 49(5)*, 735-4.

Malmgren H (1984), Habituation and associative learning in random mixtures of deterministic automata. *Göteborg Psychological Reports 15:2.* Göteborg University.

Malmgren H (1991), Learning by natural resonance. *Göteborg Psychological Reports 21:6*. Göteborg University.

Malmgren H (1997), Perceptual fulfilment and temporal sequence learning. Poster presentation, *The Brain and Self Workshop: Toward a Science of Consciousness,* Elsinore, Denmark.

Malmgren H (2002), Forced learning of graded responses. Poster presentation at the *Sixth International Conference on Cognitive and Neural Systems,* Boston, Mass.

Seung HS, How the brain keeps the eyes still. Proceedings of the National Academy of Sciences USA 93 (1996), 13339-44.