Philosophical Communications, Web Series, No. 35, pp. 163-182
Dept. of Philosophy, Göteborg University, Sweden
ISSN 1652-0459

Kvantifikator för en Dag

Essays dedicated to Dag Westerståhl on his sixtieth birthday

Presentations, re-presentations and learning

Helge Malmgren

Abstract

This paper is an argument to the effect that a certain view about mental representing, together with some very liberal constraints on the brain as a dynamic system, entails that the organism will tend to form adaptive mental representations of its environment. To show this, it will first be argued that although mental representing is a common thing indeed, representationalism, in the most important sense of that term (indirect representationalism), is false. Three different views about pictorial thinking (mental imagery, intuitive representing) are then contrasted, two of which are tied to this brand of representationalism and one of which is not. The latter view, versions of which have sometimes been presented as ”simulation” theories of imagery, is here generalised to cover all kinds of mental representation. Two models of the brain are then presented in which learning of adaptive representations follows from this theory together with certain biologically plausible constraints.

1. Mental representing and mental representations

1.1 Representationalism

Cognitive psychologists, and some philosophers, not seldom use the phrase ”to mentally represent a thing (or a fact)” as a generic verb for being in an intentional state with that same thing or fact as an intentional object. In this sense, it is not controversial that human beings often mentally represent things and facts. There are however three commonly held theses about mental representing which are much less obviously true, and which have indeed often been contested. The first one stems from Brentano and says that *every* mental phenomenon involves mental representing, or intentionality. The second one, which is not seldom (probably more often than the first) referred to as ”representationalism”, states that all intentional phenomena (all cases of mental representing) involves *mental representations*. The different meanings of this thesis and the main reason why a certain

version of it should be rejected will be the core topic of the first part of this paper. Let me also note that the combination of the two mentioned theories adds up to a third thesis, which one could call ”strong representationalism” and which is upheld by several philosophers and psychologists in the cognitivist tradition. To distinguish the three theories which have just been described, I will name them *representationalism$_1$, representationalism$_2$*, and *representationalism$_3$.* As stated, my focus will be on representationalism$_2$, the theory that all mental representing involves mental representations. What’s wrong with that thesis? I will first argue that its appeal is partly due to two possible interpretations in which it is trivially true and (therefore) theoretically uninteresting.

1.2 ”Representation” in the sense of represent*ing*

One may first note that ”mental representation” can be used as a form of the verb ”to mentally represent”. Using the term in this way, ”There is a mental representation of Pegasus going on” is just another way of saying ”Someone mentally represents Pegasus”. Then the thesis that *all intentional phenomena involve mental representation* (note the missing *”s”* at the end) surely becomes trivial. I would like to suggest that if a philosopher says ”All intentional phenomena involve mental representation*s*”, one should (if evidence to the contrary is missing) suspect that although the philosopher in question would rather like to say something non-trivial, she actually uses ”mental representations” (in the plural) as a synonym for ”occurrences of mentally representing” and so is uttering a tautology. But surely there must be more to representationalism$_2$ than this tautology.

1.2 Representing-enabling states

Secondly, it should be pointed out that any plausible theory of mental representing (intentionality) must entail that there are *representing-enabling states of the organism.* If a person first remembers a certain event and then entertains a wish that a certain other event shall occur, a certain part of the world which includes the person must be in different internal states at the two respective points of time. The reason why I use the complicated phrase ”a certain part of the world which includes the person” instead of simply ”the person” is of course that I don’t want to exclude the possibility that externalism about mental content has some truth in it. For example, immediately remembering one individual

recent sound rather than another when listening a series of exactly similar sounds might sometimes be analysable in terms of which one of the sounds was actually the most recent one (cf Malmgren 1975). However, this being said, the issue of externalism or internalism about content need not bother us further, since it is obviously true that externalism is not the *sole* truth about mental content. The difference between remembering a recent sound and thinking of Fermat's big theorem is, e.g., not solely a matter of being in different external circumstances. Hence there must also be internal states of the organism upon which an essential part of this difference depends.

We will henceforth refer to these states as *mentally-representing enabling states*, or *representing-enabling states* for short, bypassing the issue whether or not they *alone* determine content. Note that "mentally" in the full phrase "mentally-representing enabling states" signifies that the representing which the states in question enable is mental, not that the states themselves are necessarily mental. As far as our definition of these states go, they may be states of the brain, and the intentional content may be determined by these brain states in the sense that it is caused by them. Now, part of the appeal of representationalism$_2$ might stem from the possibility of interpreting it as the rather obviously true thesis that *intentionality must involve mentally-representing enabling states*. But again: surely there must be more to representationalism$_2$ than this.

1.3 *Mental* representing-enabling states

What about interpreting of representationalism$_2$ to mean that there must be *mental* representing-enabling states for intentionality to be possible? In this interpretation, representationalism$_2$ results from taking what I just stated to be a necessary component of any plausible theory of mental representing (intentionality) and adding the condition that the states of the organism which (wholly or partly) determine intentional content are themselves *mental*.

The last occurrence of "mental" in the last sentence can be given many possible meanings more or less independently of what one puts into the occurrence of the same term in the phrase "mental representing". This entails both that the resulting theory we are talking about

is actually a family of theories, and that it need not be trivially true. One theory in this family is the Husserlian one, according to which intentional content is constituted in acts of consciousness which are themselves conscious because they are *experienced* (erlebt) (but not thought of). A quite different possible theory in the same family says that content is determined by wholly unconscious, but still mental, cognitive processes. In the following I will treat all the family members as versions of one and the same generic theory, and I will *not* argue that there is no true theory of this kind.

1.4 Indirect representationalism

It must be noted that the last-mentioned generic theory, as such, does not give any support to certain ideas which are central in much contemporary representationalism. I am referring to the fact that properties such as systematicity and compositionality are regularly ascribed to mental representations. The bare theory that intentionality (mental representing) requires *mental* representing-enabling states may not be trivially true, but it does not entail *anything* about the nature of the states in question, except that they are mental.

A stronger version of representationalism$_2$ results if one models mental representations, in the sense of the generic theory, on the use of *public* representational systems. Public representations, such as linguistic symbols and ordinary pictures, depend for having their meaning on their sometimes being *perceived* by human beings, who when perceiving them also apprehend something else *through* them. Many representationalists postulate mental entitities – mental representations – which are similarly perceived or otherwise apprehended in some relevant way by the subject, and they believe that the content (intentional object) of a mental representation is being apprehended *through* this first apprehension of the representation itself. In other words, intentionality or mental representing is an essentially *indirect* affair. I will henceforth reserve the term ”representationalism” for theories which conceive of mental representing in this way, but to avoid confusions I will also refer to such theories as brands of *indirect representationalism.* Note that in order for indirect representationalism not to collapse into the generic theory that intentionality is due to mental representing-enabling states, the way in which these states are supposed to be *apprehended* must be something over and above the circumstances which make them

mental. And a Husserlian theorist can avoid being committed to indirect representationalism through denying that being conscious in the sense of *erlebt* entails being *apprehended.*

A good example of an indirectly representational theory is the age-old idea that mental imagery should be analysed as the perception of ”inner pictures”. In the philosophy of perception, indirect representationalism can be exemplified by (most of) the well-known sense-data theories. These say that perceiving a physical object involves sensuously apprehending a sense-datum and taking it as a sign of an external object. Finally, the linguistically oriented representationalists of today are as a rule committed to the hypothesis of indirect mental representing. When they do not state this hypothesis explicitly, one may sometimes infer its presence from indirect evidence. The formulation ”symbol manipulation” is especially revealing. Why should one speak of intentionality as involving the *manipulation* of symbols, if one did not believe that the symbols in question are somehow *apprehended* by the subject? From the much less controversial thesis that mental representing requires mental representing-enabling states it does not seem to follow that anything at all is being ”manipulated” when intentionality is at work. Some additional premise is obviously used here, and it is my conviction that the mentioned analogy with the use of public representations is the missing link.

One may of course argue for the systematicity and compositionality of representing-enabling states without supposing that these states are being apprehended in any way by the subject whose mind or brain they are states of. However, I believe that the idea of such language-like properties of mental representations borrows much of its credibility from the plausible thought that these properties make *public* representations easy to manipulate by a human being who apprehends these representations. If one wants to argue that systematicity and compositionality pertain to a system of representing-enabling states which are *not* apprehended and manipulated by a subject, one will have to show for a start that their universal occurrence is a highly desirable property from a brain-theoretical and evolutionary perspective. And that is a large order.

1.5 The simple regress argument

It is of course important for the proponent of indirect representationalism that the sense in which she supposes that a mental representation must be ”apprehended” does not itself involve *mentally representing the representation*. Else, an infinite regress will immediately ensue. Some indirectly representational theories may have been formulated with so little sophistication that they are vulnerable to this ”simple regress argument”, as I will call it. However, most theories avoid it by emphasizing that the apprehension of a mental representation is a special kind of process which cannot be analyzed in terms of mentally representing the representation. I will not here go into the problem of explicating the nature of the apprehension in question. To be sure, this may turn out to be a very difficult task considering that for indirect representationalism, apprehending an object *through* a mental representation must involve something over and above just *being in a mental representing-enabling state*.

1.6 The problem of meaning-giving

The point against indirect representationalism which I will now make should also be familiar from the contemporary discussion, but it can be worthwhile to repeat it since it may be confused with other arguments, including the simple regress argument discussed above. It can be formulated as ”the problem of meaning-giving”, and also leads to the conclusion that indirect representationalism entails an infinite regress.

That a public symbol (such as ”green”) has a meaning entails that subjects who perceive the symbol regularly apprehend that meaning through perceiving the symbol. If indirect representationalism is correct, mentally representing a meaning similarly involves apprehending the meaning through apprehending a mental representation. But just as perceiving the physical symbol is not in itself sufficient for taking it as having a certain meaning (since it can be perceived but taken as having another meaning), apprehending a mental representation should not in itself be sufficient for taking it as having the meaning it has. In the case of the public symbol, it is tempting to say that what is added to the bare perception of the physical symbol is an act of mentally representing, in which a meaning is ascribed to the symbol. However, this solution obviously does not work for mental

representations as conceived by the indirect representationalist, since it would mean that yet another mental representation is involved, and so would lead directly to an infinite regress. How, then, is meaning given to the mental representation?

An especially well-known version of the problem of meaning-giving is the Wittgensteinian argument against the ”inner picture” theory of intentionality. It starts with pointing out that public pictures do not always work by means of exact similarity, and continues by noting that perceiving the meaning of such a picture therefore needs something like an act of interpretation. Hence if mental representing consists in an analogous process in which mental pictures are apprehended, these pictures will also need interpretations, which means that another infinite regress is started.

It is important to note that the essence of this Wittgensteinian argument is *not* the Berkeleyan point that exact similarity is not particularly well suited for representational purposes. Neither is the point that of the simple regress argument, which presupposes that apprehension of a representation entails representing the *representation*. Instead, the core of the argument from meaning-giving is that indirect mental representing seems to need something more than the apprehending of a representation to work – namely, mentally representing its *meaning*. This in turn entails the existence of another mental representation, and so on.

To be sure, supposing that *some* mental representing is done by means of indirect mental representations does not lead to an infinite regress. Maybe it happens sometimes that we do ”interpret” something that can be called a ”mental symbol” or a ”mental picture”. But it is not possible that *all* mental representing works this way, and an *analysis* of mental representing in terms of indirect mental representations cannot therefore be true.

Now, we all know that some modern representationalists have sought a way out of this threatening regress by postulating that the mental representations get their meaning not from a mental process of meaning-giving, but from some kind of *natural* (causal, or informational) relation between the representations and the world. Causal theories of

meaning surely have problems of their own, but let us suppose that they can be overcome. The question remains, in what sense do we now have *mental representations* in a sense over and above that of *mental representing-enabling states*? Suppose, as the causal theories of the meaning of mental symbols assert, that certain states of the mind are informationally coupled to certain objects or facts in the world, and that this coupling underlies our talk about ”mentally representing” the latter. What reasons do we have for thinking that these representing-enabling states are ever apprehended by the mind so as to be accessible for manipulation and combination? I would say, none. And of course the indirect representationalist does not convince us on *this* matter if he instead argues that the mentally-representing enabling states in question are physical rather than mental.

My general conclusion is, therefore, that although mentally representing is a common and undisputable phenomenon, and although such representing surely needs representing-enabling internal states, and possibly needs mental such states, mental representations in the strong indirect sense required by some contemporary representationalists are *not* involved in all mental representing.

2. The nature of mental imagery

2.1 Three theories of mental imagery

As uncontroversial as it is that there is such a thing as mental representing, it should be that there is such a thing as mental imagery (pictorial thinking; in German: *Anschauliches Vorstellen*, in Kantian contexts sometimes translated as ”intuition” or ”intuitive representing”). If there seems to be a controversy about the existence of mental imagery, this is usually due to the fact that the concept of mental imagery has been laden with a specific theory about its nature. Which theories are there, then, about this nature?

There is of course the already mentioned (and often scorned) theory that mental imagery consists in the apprehension of certain picture-like entities, *mental images*, that represent their objects by means of similarity. Apart from the general argument against indirect representationalism outlined above, this view also suffers from problems related to the choice of similarity as the representing relation. It shares many of the latter problems with

analogous versions of the so-called ”representative theory of perception”. First, of course, we have the Berkeleyan argument about abstract general ideas. Then, how can a green quale, whether occurring in a perception or as a component of mental imagery, represent anything in the world *by similarity* if the nature of objective (represented) greenness is very different from that of the quale? This argument easily leads to the division of perceived properties into ”primary” and ”secondary” ones, where the primary ones are the properties (such as shape and relative position) which can plausibly be represented through their similarity with mental items. It is no coincident that some contemporary authors (notably Kosslyn) who still argue that mental images represent by similarity tend to exemplify their thesis with the representation of spatial structures.

Another, more radical way out of the problem is to argue that what mental images depict by similarity is not the external objects, but our perceptual states. The extra cost of this solution (over and above the problems which it still shares with the first theory) is of course that the reference to the world is left unexplained, not only for mental images, but even more so for perception. So, this theory offers a natural way to subjectivism in epistemology.

There is however a third theory about mental imagery which has the advantages of the last-mentioned solution without sharing its most serious drawbacks. It is the thesis that the similarity of mental imagery to perception is indeed essential to the working of the former, but not because mental images *represent* perceptual states but because mental imagery *has the same representing-enabling properties as* perception. In other words, mental imagery is perception-like in its intentionality.

Obviously, this third theory does not entail that mental imagery involves mental *images* in the sense of mental entities which are somehow apprehended by the subject. Neither does it entail any other specific thesis about how mental imagery represents the world, except that it does it in the same way as perception. It is compatible with almost any theory about the nature of perceptual intentionality. Of course this can be seen as a drawback. However, since the present theory avoids the very serious problems pertaining to the indirect representationalist interpretations of mental imagery, and since many interesting points

about mental imagery and mental representing in general can be made with it as the starting point, I think it is much to be preferred if one has to choose between the mentioned candidate theories. It will now be described in some detail.

2.2 Representation as re-presentation

A good illustration of how mental imagery can fulfil some of the functions of perception is given by what happens when you light up a dark room for just a second. As a rule, you are then able to navigate through the room for several seconds before your vivid, intuitive representing of the room comes to a (more or less gradual) stop. The same mechanism is at work when you are out walking but direct most of your attention to other things than the path. It suffices to cast a brief glance at the path now and then in order to walk safely. One interesting point to be noted is that the intuition (i.e., the intuitive representing) which steers your steps between these glances is *dynamic* in the sense that it tends to update itself continually. In the walking situation this feature is of course essential to the ability of the intuition to guide your steps for more than a single moment. I will give an explanation of this important phenomenon below.

In the situations just described, it is obvious that the mental imagery works as a functional *substitute* for perception in a certain field of intentional tasks, viz., for navigation. It is of course a long step from this observation to the full-blown theory that all mental representing (except, of course, perception itself) can be analyzed in terms of such substitution of intentional functions. I will not try to prove this general thesis, but one aim of my paper is to show that certain things can be explained in a natural way if it is true.

Before that, let me note as a kind of indirect argument for the theory that it does not entail that mental representing is ever *exactly* like perceiving. Even pictorial thinking is usually much different from perceptual experiences, a fact that Hume may have been the first to find philosophically interesting. In the example with the brief lighting-up of a room, you *may* have a fairly vivid and perception-like positive visual after-image after shutting off the light, but the visual after-image may also be weak and your experience dominated by a more kinaesthetically toned ”feeling” for where the obstacles, openings and other affording

structures are located. This feeling still usually suffices for the mental imagery to fulfil its navigatory function. So, a case of mental representing may be sufficiently *functionally similar* to perceiving even if there is *no obvious phenomenological similarity* between the respective states. One should bear this remark in mind when discussing the generalised form of the thesis under discussion, so as to avoid objections based on the obvious phenomenological dissimilarities between perceiving and thinking (e.g.).

The thesis that all mental representing except perception itself (in the following, the latter qualification will usually be left implicit) works by its intentional-functional similarity to perception is sometimes (e.g., Hesslow 2002) referred to as a *simulation theory*. This designation is well found, but sometimes I prefer to use *representation as re-presentation* as a brief descriptive name of the theory. One reason for this is that the term ”re-presentation” hints at another element which can be combined with the basic tenets of the theory in a very natural way. I am referring to the Humean thesis that ideas are ”copies” of impressions not only in the sense that ideas are similar to impressions, but also in that the former are *derived from* the latter. In the example with the dark room it is fairly obvious that the representation that you have after the light has gone out is, to a large part at least, a re-presentation of the perception you had while the room was lit up. In the walking example, the connection is looser (since the representations are dynamic) but still the successive representings are obviously strongly dependent on the successive perceptions. The third part of this paper can actually be seen as an explication, in modern system-theoretic and learning-theoretic terms, of Hume’s doctrine of the derivation of ideas from impressions.

3. How to explain learning

3.1 Perceiving, representing and feedback

In the present section, I will try to briefly describe perceiving and mental representing from a brain-theoretical perspective. Considerations about the relations between brain and mind will be wholly left out; it is just assumed that to every mental state (type) there corresponds an underlying brain state (type). The reader will also notice that the term ”information processing” is not used in my description. This is because the phrase has been used so much

by representationalist philosophers and psychologists that it is difficult to free it from its homuncular connotations.

Our sensory receptors are continually influenced by energies in the environment. Their response patterns are transformed by neural mechanisms, and the transformed patterns in turn influence both other neural structures and devices outside the brain, such as muscles and glands. At some fairly high level of transformation, some of the neural patterns form the basis of what we call ”perception” (or ”perceiving”). Like other neural patterns, the perceptual ones are to a large part fed back to the brain, but they are also used to directly shape the responses of extra-brain structures. This description of the brain in general and perception in particular can be summarised by saying that *the instantaneous perceptual neural pattern can be regarded as an output signal of the brain-system, a signal which is also continually fed back to the system*.

Note that this formula does not entail that each and every aspect of the instantaneous perceptual pattern has a specific effect on either the brain or the extra-brain body; in other words, it allows for the possibility that only part of the perceptual signal is actually used in either of its functions. With this qualification, the formula should not be controversial, although the way of looking at perception which it reflects may not be the most common way. The emphasis on *output and feedback* will come as a slight surprise to anybody who is used to regard perception either as an input signal, or as a pure end product of the brain, or as just a mediating link between ”lower” brain processes and the responses of the whole organism. To help familiarise the reader with the present perspective, I want to point out that the feedback from perception underlies much of what we usually call *memory*. Both episodic short-term memories of recently undergone perceptual experiences, explicit ”semantic” knowledge based on generalisation from earlier observations, and direct modifications of perceptual responses based on earlier perceivings (perceptual memory) surely depend on the earlier perceptual patterns having been thus fed back to the brain.

From the same system-theoretical point of view, a main difference between perceiving and (other) mental representing is that the latter is more or less *decoupled* from the environment.

In the by now well-known paradigmatic case where we use mental imagery to navigate in a dark room, this decoupling is effected by the absence of relevant current visual stimuli. In other situations, we close our eyes to think better, or just concentrate on our thoughts in order not to be disturbed by current external events. Using a gross simplification one can say that *while perceiving is mainly externally steered, mentally representing is mainly internally steered*. Note that to a large extent, ”internally steered” in this statement should be taken to stand for *steered, through feedback, by earlier input*. Hence the formula should not be taken as implying that external factors do not influence mental imagery and other mental representing, only that the most important kind of such influence is *delayed* and not immediate.

Now, what the thesis of mental representing as a simulation of perceiving says is that the just described difference between perceiving and (other kinds of) mentally representing is the *only* important functional difference between them. *Both perceptions and mental representings are output patterns of the brain, which are fed back to the brain as well as used in the immediate shaping of bodily responses. The main difference between them lies in the nature of their input dependence.* Below, I will show that these seemingly innocuous propositions, together with a few simple constraints on the brain as a system, entail that adaptive representing states will tend to be formed in the brain. In this derivation, an essential part will be played by the assumption that the neural patterns underlying mental representing are, like our perceptual patterns, fed back to the brain. That this assumption is true should also be fairly obvious since we do remember a lot of our imagery and thoughts, not only our perceptions.

Before we go on, note that one plausible associate of functional similarity is similarity of structural basis. In other words, a straightforward way of implementing the functional similarity between perception and representation is to base them on similar patterns in the same neural structure. Several kinds of empirical data indeed support the idea that perception and mental imagery share the same neural structures to a considerable extent. This is usually (and correctly) taken to be evidence for the simulation thesis. It will also be a consequence of the model which is outlined below that such sharing of structures occurs.

3.2 Self-organisation of the perception simulator

The essence of my argument can be summarised in the following way. Imagine that for some reason, the brain-system alternates between having normal input from the environment and being informationally shut off from the environment. During these respective conditions, the brain will produce outputs which we call ”perception” and ”representation”, respectively. In both conditions, the output will be fed back to the brain. Now, if the output during epochs of representing is similar to the output during perceiving, the brain will probably react in a similar way to the feedback during the two kinds of epochs, while if it is dissimilar, it will probably react dissimilarly. This means that everything else being equal, *producing similar outputs during representing as during perceiving will be a way for the brain to uphold stability.* In the long run, the brain will therefore tend to go to a state in which such similarity of output is being upheld.

The following subsections will be devoted to three more specific models where this fairly vague and general argument is replaced by exact counterparts.

3.2.1 The simple finite model: alternating between constant input and ”no input”

In this subsection and the two that follows, the memory system of the brain will be modelled as a *finite deterministic system* with input and output, where the output is also fed back to the system. In order to try to catch some interesting general tendencies of such systems, let us look at the *randomly composed* system. This is a stochastic entity where all items in the transition table are randomly chosen, with uniform probabilities, among the states of the system. Its outputs, the set of which is supposed to coincide with the set of perceptual states, are also randomly assigned to its states.

The randomly composed system should not be confused with the uniformly probabilistic system which at each moment chooses with equal probabilities which state to go to. Instead, the randomly composed system should be thought of as a big collection of systems, each of which follows its *own* laws in a *fully deterministic* way. Hence when, below, I speak of the

probability that the randomly composed system will perform in a certain way, this refers to the proportion of all possible finite deterministic systems which will perform in this way.

In our first model of the brain, it receives a certain environmental input A alternatingly with being shut of from the external world. It is here presupposed that perception works in a simple transduction-like manner, so that from the same input A the same perception B is always produced as the brain's output. In the periods of being "shut off", it is supposed that the output is instead wholly determined by a separate system of the brain, which we designate as its "memory module". The question marks symbolise the output of the memory module when the system is thus "internally steered".

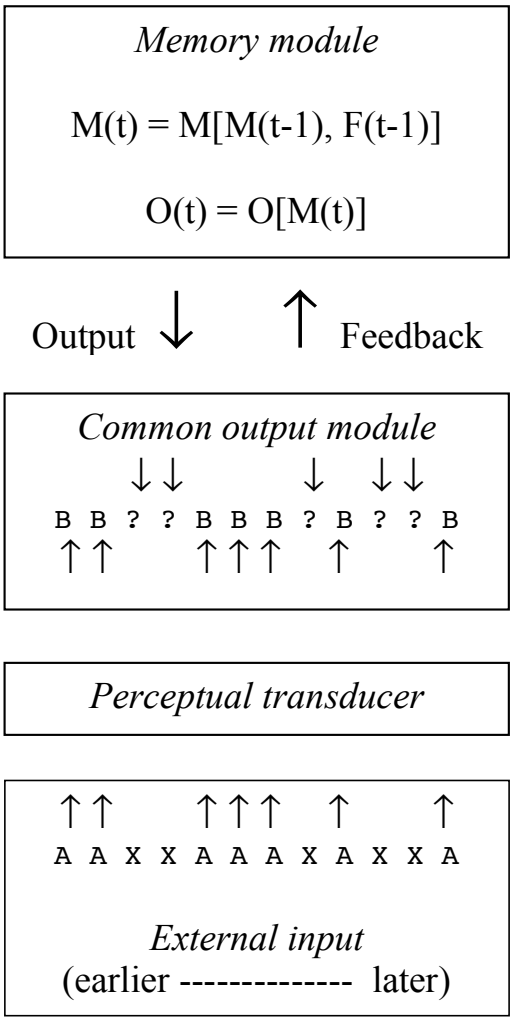

Figure 1. Learning about a constant environment

To see that systems which are organised according to *Figure 1* will tend to produce the same output when the input is shut off as when it is not, suppose that the memory module is a randomly composed system, and first consider the case where it has reached an point attractor – a stable state S – under the constant perceptual input B. There is then a non-zero probability that the transition table of the memory module says that it shall produce output B when it is in state S. If so, the memory module will stay in its attractor during the period of environmental shut-off, since its input will stay the same (through the feedback loop) as when it perceived the world. If it produces some other output than B, this difference will make a difference to its input during the shut-off epoch, and there is a finite probability that it will go to another state S'. This state may in turn produce B as output, in which case the system is again stable. Hence with time, the probability of output B during periods of shut-off will rise; in other words, more and more systems in the big collection of systems will stably produce output B.

The above argument is formally incomplete since there are other possibilities for the system to become stable. For example, the memory module need not be sensitive to a change in feedback between B and, say, C; in which case it may end up giving C as a constant response during the periods of shut-off even if the perceptual response is B. However, for no other possible output than B is there a specific mechanism of stabilisation corresponding to one described, which means that the general conclusion still holds. Similar considerations hold for the situations where the memory module is not in a point attractor.

3.2.2 The dynamic finite model: learning a sequence of inputs

Learning to substitute for a constant perception has some uses, for example when one finds oneself in a dark but non-changing room. What about dynamic environments? Somewhat surprisingly, our random finite system will also tend to learn to reproduce *repeating perceptual sequences*. The essence of the proof of this proposition is the observation that if the memory module actually by itself produces the same sequence as the perceptual one, *and in phase with it*, the system will stay stable when it hears the external sequence again (the auditory modality naturally comes to one's mind here). This is, again, due to the

feedback from the common medium of perceiving and representing. As in the previous case, other possibilities of stabilisation will not outweigh the mentioned one. For details, see Malmgren (1996, 1997).

3.2.3 The finite Pavlovian model: learning to substitute for unconditioned stimuli

To be able to describe our last finite example in an economical way, I will use the same letter to designate an external input and the corresponding output of the perceptual transducer. Suppose, then, that the brain which illustrated in *Fig. 1* is exposed to the following environmental conditions. First, it receives a number of sequences ABC and in between these sequences a constant ”background” stimulus D. After a number of ”trials” with ABC, it receives just AB, after which the input is immediately blocked. Which output does the memory module produce? In simulations (Malmgren 1991) it has be shown (i) that the most common output will be C, (ii) that the frequency of output C on input AB will rise with rising number of previous presentations of ABC, and (iii) that the ”conditioned response” C is *specific to the sequence* AB of ”conditioning stimuli” since, for example, BB does not produce nearly as many C responses as AB does. All these features are typical of classical (Pavlovian) conditioning.

How can this performance be explained? The main case to consider is now the one where the memory module has arrived at a stable state S such that (i) S is stable under the background input D and the sequence ABC, and (ii) *if* the system receives input (or feedback) AB while in S, it gives C as its final output. It is obvious that such a system must be stable under the experimental conditions, while a system which fulfils (i) but not (ii) runs a fair risk of being destabilised – again because it may be sensitive, through the feedback, to the difference between percept and representation.

3.2.4 A continuous model: the perfect learning system

The finite brains considered above are bad learners since they rely on a random search for a stabilising solution. The opposite property is shown by those systems which have a *learning continuous attractor*. In the two-dimensional case, such a system is defined as follows (for more details, see Malmgren 2002):

For a certain region of state space, holding x constant at any value x_0 causes y to approach a value y_0, such that holding y constant at y_0 causes x to approach x_0.

You should here think of x as a perceptual state and of y as a state of a memory module. In *Figure 2* below, holding x constant is referred to as *presentation*, the process which it initiates in the system as *storage,* and its final y-wise result as a *representation*. Holding y constant, i.e., letting the representation control the system state and re-create the originating perception, is called *retrieval.* The green line in the figure is the continuous attractor, or set of contiguous point attractors, which characterises the system's asymptotic behaviour.

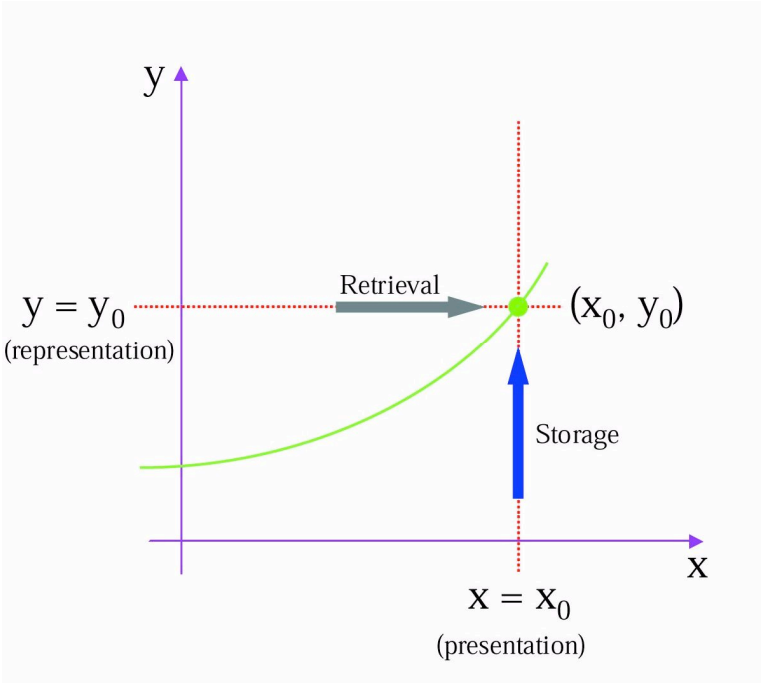

Figure 2. A learning continuous attractor

It can be shown (Malmgren 2001, 2002) that systems whose time derivatives fulfil certain simple qualitative conditions will have such continuous attractors. These systems will be perfect learners in the sense that if exposed to a certain presentation, they will in the long run *always* produce a representation which in turn can re-create the presentation if allowed to steer the system.

So, Hume's dream has been fulfilled by the learning continuous attractor. With this, I have completed my demonstration that the simulation theory of mental representing lends itself very naturally to simple explanations of learning.

Helge Malmgren
Department of Philosophy
Göteborg University
Box 200, SE-405 30 Göteborg, Sweden
helge.malmgren@filosofi.gu.se

References

I have omitted a host of possible references to the contemporary debate about mental representations.

Hesslow G, "Conscious thought as simulation of behaviour and perception." *Trends in Cognitive Science 6:6* (2002), 242-247

Malmgren H, "Internal relations in the analysis of consciousness." *Theoria 41* (1975), 61-83

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

Malmgren H, "Perceptual expectations and the learning of temporal sequences". *Philosophical Communications, Red Series 35,* 1996

Malmgren H, "Perceptual fulfilment and temporal sequence learning". Poster presentation at *The Brain and Self Workshop: Toward a Science of Consciousness.* Elsinore, Denmark 1997. Available on the Internet as http://www.phil.gu.se/posters/resonance

Malmgren H, "Representations can re-present. Notes on unsupervised online learning in sleep-wake systems with continuous attractors." Oral presentation at the *Fifth International Conference on Cognitive and Neural Systems*, Boston 2001. Accessible on the Internet as http://www.phil.gu.se/posters/HMBB.htm

Malmgren H, "Forced learning of graded responses." Poster presentation at the *Sixth International Conference on Cognitive and Neural Systems*, Boston 2002. Accessible on the Internet as http://www.phil.gu.se/posters/hmgraded.pdf