Vehicle 12

Trains of Thought

At this stage, if you want to be critical, it is easy for you to maintain that up to now you have not discovered anything in our vehicles that goes beyond ordinary learning. True, these creatures seemed to become more and more able to deal with the adversities of their environment, not only by a process of Darwinian selection but also by active assimilation of information from the world. But thinking is different. It is a process that can go on for a long time, as everyone who has done some conscious thinking knows. Thinking can be observed in other people as well, when we get verbal or nonverbal evidence for a succession of mental states that are guided by some criterion of plausibility or logic—mental states that reflect the exploration of various blind alleys and eventual arrival at a result. Sometimes we seem to notice such mental operation even in a monkey or in a dog. But not yet in a vehicle.

The possibility of sustaining long successions of distinct brain states for the purpose of exploring knowledge already incorporated in the brain is what we will introduce in a new brand of vehicle, which we will call Vehicle 12.

First a remark on pathology. All the later vehicles, beginning with type 7, are in constant danger of running into a condition

quite analogous to epilepsy (which is also one of the most common forms of derangement of animal brains). The strengthening of the connections between the elements of the brain, which is at the basis of associative learning, embodies the danger of reciprocal activation beyond control. In a population of elements in which excitatory connections abound, if the number of active elements reaches a certain critical level, chances are the remaining ones will also become activated. These elements, in turn, keep the first set active. A maximal condition of activity is then established and maintained until the supply of energy is exhausted. This maximal activation makes no sense in terms of the information ordinarily handled by the brain, which is keyed to patterns of partial activation of the elements. Necessarily the result is disorderly, ineffective behavior. There are various ways of dealing with this danger, and I propose the following for our vehicles.

Let every threshold device in the vehicle's brain be touched by a special wire through which we can control its threshold. If we set the thresholds high, the threshold devices will become active only when they are very strongly activated by the input they receive from other threshold devices or from the sensors. For a lower threshold, less input will suffice. So if we watch the operation of the brain—and in particular the total amount of activity in it—we can always prevent an attack of epilepsy by raising all the thresholds. If there is not much activity, we can lower all the thresholds and thereby encourage the circulation of activity through the brain. It is of course quite easy to let this happen automatically. All we need (figure 18) is a box that receives as its input the number of active brain elements at that moment and calculates appropriate thresholds, which it then sets for the whole brain. In real life, the input for this threshold control device might be the rate of change of the number of active elements, in order to give it an opportunity to foresee the catastrophic explosion of activity before it happens. But

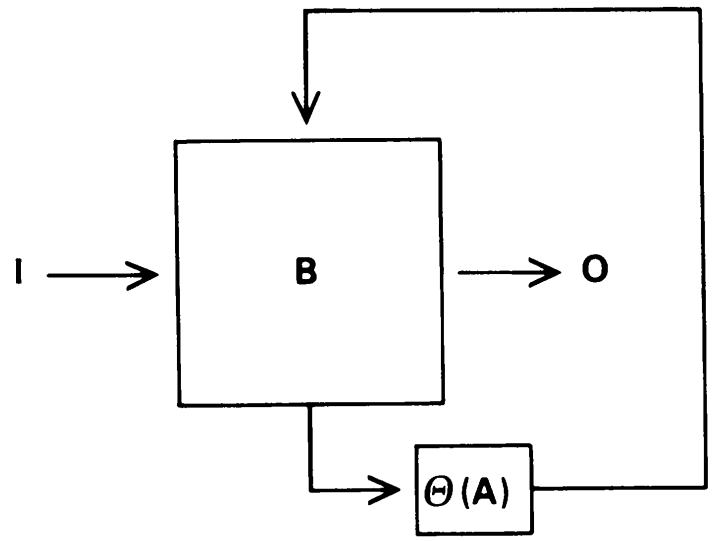

Figure 18

B is the brain, which receives input I and elaborates an output O. At the same time it signals the level of activity A in its interior to a special box that calculates appropriate thresholds Θ for the elements in B.

for purposes of illustration it will suffice if the threshold control device works just on the amount of activity in the brain.

The effect of this global negative feedback on the activity of a vehicle's brain is illustrated in figure 19, which shows the number of active elements as a function of the number of active elements a moment earlier. When the activity is low, it will again be low at the next moment. (For very low excitation, there may even be a tendency for the activity to die out, since a minimum density of active elements in the brain is required to activate the next set of elements, but this is not shown in figure 19.) For very high levels of excitation—that is, for a very large number of active elements—we may imagine that the thresholds are immediately set so high that the activity will drop to a very low level at the next moment. Intermediate levels of activity will lead to maximum activity at the next moment (see the middle part of the curve in figure 19). Later on we will come back to this curve, which has interesting philosophical implications. First let us watch the operation of a brain that con-

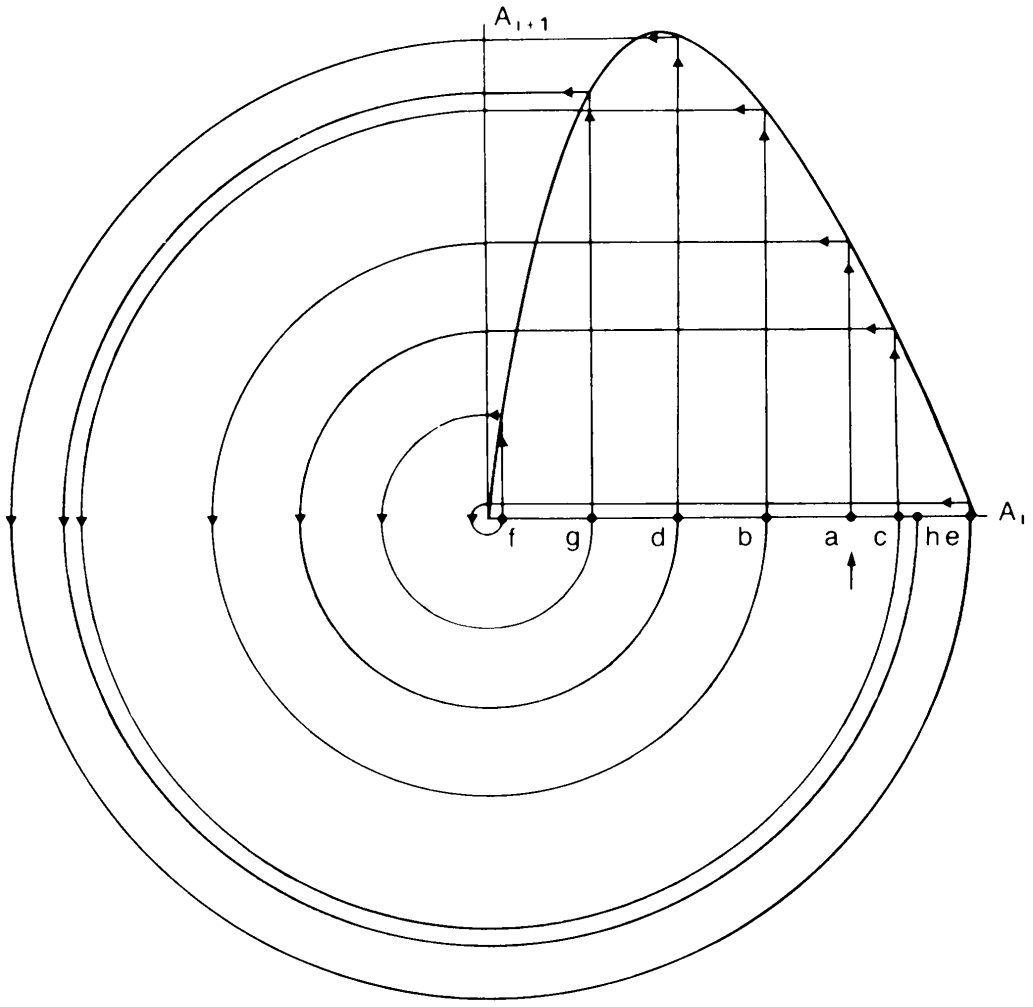

Figure 19

The function describing the next number of active elements A_{t+1} given the present number of active elements A_t. It can be seen by iteration (follow the lines starting from the arrow near a) that the states of a brain controlled by such a rule are quite unpredictable.

tains many learned associative connections while it is being controlled by the feedback of a threshold control device.

We have already noticed that the vehicle's brain has a tendency to explode into fits of activity because of the abundance of reciprocal activation between its elements, a situation reminiscent of the chain reaction in a block of uranium. But most of these explosions, if everything works out the way we have planned, should take place within limited groups of elements that are tied together by particu-

larly strong associative connections. Such sets of elements arise as "concepts" representing things or events that have often presented themselves in the environment.

Let one such thing appear in the sensory space of Vehicle 12. The explosion of activity will happen in the corresponding set of threshold devices responsible for that concept. This implies an increase of the number of active elements in the brain, and the threshold control device will immediately react to it by raising all the thresholds. A moment later many elements that were previously active will be silent. But the elements pertaining to the concept in question are likely to stay active. This is because the strong reciprocal connections within the set, once activated, guarantee a very high level of excitation for each element of the set. This level is so high that the activity of the elements may survive the raising of the thresholds. Thus the first interesting effect of our recent innovation is the focusing of individual concepts—of patterns that have their own internal consistency—at the expense of background activity. We greatly appreciate this effect in a well-functioning human brain, where it is often called the FOCUSING OF ATTENTION.

But there is more. You remember that we have installed not only Mnemotrix wire for concept formation but also Ergotrix wire, which represents within the brain the relation of temporal succession, of consequence or causality. Thus the elements now active in the lone surviving concept after the automatic raising of the thresholds also have some Ergotrix wires attached to them. These Ergotrix wires lead to the elements that have often been activated after the concept in question, the consequences of the active concept, so to speak. Obviously, there will be more than one possible next step for all but the most determined situations.

So we must ask ourselves how the vehicle's brain finds the concept that follows the one it presently holds. The choice, it turns out, is quite automatic. Among all the elements activated by the present concept through the Ergotrix wires, there will be some groups

strongly connected by Mnemotrix wires because they again form concepts. These groups will of course ignite with particular alacrity because the internal connections within each group will provide an explosive kick to the activation from external sources, that is, from the active concept. Now you can see what will happen. The threshold control, alarmed by all this growth of activity, will quickly raise the thresholds, smothering most of the activity and leaving only the most resistant group of elements activated. As we have already seen, this will be the group with the strongest reciprocal connections. In terms of concepts we may put it this way: the next concept, among all the concepts that are possible consequences of the present one, will be the most consistent or familiar one—the one most strongly established by experience.

Note that with all these budding and growing explosions the thresholds have been raised above the level at which they were set for the previous concept. It is therefore very likely that the previous concept will be extinguished. So the system will not swing back into its former condition but will end up with a different concept. This new concept will have its own consequences embodied in Ergotrix wires. And these will again materialize in a new concept by way of the sequence of events that we have just described. The process will continue as long as you wish or as long as the chain of concepts does not lead back to the concept from which it started.

The upshot is something very much akin to thinking, to that process so familiar to our introspection, where images appear in succession according to rules reflecting the relations between the things they stand for. This process goes on in our minds when we try to figure out the best way to get from one point to another in a familiar city by letting our imagination produce successions of street corners (or other landmarks) whose relations of geographical proximity we have experienced. It is also one of the tricks we use to determine the consequences of possible moves in a game of chess, or the consequences of some statement in a discussion. This chain-

ing of internal states is exactly what we planned to introduce into the brain of Vehicle 12 to make its meditations look more lifelike, more like our own, not only in the time they take but also in the unforeseen routes they can follow.

There is an important property that the brain of Vehicle 12 shares with the brains of our fellow men. Consider again the curve of figure 19, which shows the number of active elements as a function of the number of active elements a moment earlier. The exact shape of the curve is not very important, as long as it has a maximum and cuts the diagonal ($A_i = A_{i+1}$). Start with a certain value *a* on the abscissa and find the ordinate of the next value *b* on the curve. Put that value *b* again on the abscissa and find *c*, and so on. You will be surprised to find that the succession of values *a, b, c . . .* does not seem to follow any rules and is in general quite unpredictable. Now you will remember that figure 18 describes the effect of threshold control on the activity of the brain of Vehicle 12. We may take *a, b, c . . .* as the number of active elements in the brain in successive moments of time. If there are very few elements, the succession will by necessity become repetitious after a short while. But for a fairly large brain the succession will be truly unpredictable to an observer, for any practicable stretch of time.

I hope you realize what this means. If you could observe the inner workings of the vehicle's brain, say, by watching light bulbs connected to the threshold devices, and these light bulbs lit up every time the corresponding element became active, you could not even predict how many lights would light up in the next moment, let alone what kind of pattern they would form. (For any given number there are of course many constellations with that number of active elements!) At this point we should again invite our philosophers to comment.

I would claim that this is proof of FREE WILL in Vehicle 12. For I know of only one way of denying the power of decision to a creature—and that is to predict at any moment what it will do in the

future. A fully determined brain should be predictable when we are informed about its mechanism. In the case of Vehicle 12, we know the mechanism, but all we can prove is that we will not be able to foresee its behavior. Thus it is not determined, at least to a human observer.

I know what the philosophers will reply. They will say that although this may look like free will, in fact it is not. What they have in mind when they use that term is the real power of decision, a force outside any mechanical explanation, an agent that is actually destroyed by the very attempt to put it into a physical frame.

To which I answer: whoever made animals and men may have been satisfied, like myself, a creator of vehicles, with something that for all intents and purposes looks like free will to anyone who deals with his creatures. This at least rules out the possibility of petty exploitation of individuals by means of observation and prediction of their behavior. Furthermore, the individuals will themselves be unable to predict quite what will happen in their brains in the next moment. No doubt this will add to their pride, and they will derive from this the feeling that their actions are without causal determination.