A THEORY OF INTELLIGENCE

by: P. J. vanHeerden
POLAROID RESEARCH LABORATORIES
750 MAIN STREET, CAMBRIDGE, MASSACHUSETTS
02139

The basis of this theory of intelligence, in man or animal, is a model. The model is a computer, "the brain", with input channels and output channels. Shannon, in his theory of information[1] has shown that all information channels can be presented in the form of binary time series, sequences of ones and zeros only, as for instance 101101..... This is so because any record of events or things in the real world-be it photographs, television tapes, audio tapes, written or printed material, or any other imaginable record contains only a finite amount of information. Therefore, the model of intelligence can be a finite state digital computer, having only binary time series as inputs and outputs.

A further choice will be made in the model. It will be assumed that the program, that is the set of instructions given to the computer to make it into a model of intelligence, is independent of the size of the computer. This means that the model for intelligence in man, mouse, octopus or ant is the same in principle, although the amount of information processing needed to model human intelligence is clearly much larger than in the model of an ant.

It is almost silently implied that the program for the computer, to make it into a model of intelligence, is completely independent of the "actual meaning" of the information being processed. In a binary computer, the machine can only see the difference between the symbol 1 and the symbol 0, and nothing more.

If a man is hungry he will try to satisfy his hunger by eating. That is intelligent. The general principle of intelligence will be considered the generalization of that statement: "To be intelligent is to try to satisfy one's personal drives". However, one could not understand man, or social animals, without assuming the existence of several independent social drives-as for instance the parental instincts-which may very well overrule the drive of hunger in specific situations. If one is hungry, it may still be more satisfying to feed one's child than to eat one self. Man's spectrum of social drives has been very well represented for instance in the work of the social psychologist William MacDougall[2])

The drives will be represented in the model as a number of input time series to the computer. To satisfy one's drives one has to act, operate on the outside world, in some way. One may for instance ask where to find a restaurant, one may walk to it, one may pick up a menu. It will be assumed that intelligence results in commands to the muscles, as to the tongue in speaking, the legs in walking, the hand in holding the menu. These commands will be represented in the model as the output binary time series of the computer. There are, besides the drives, a second type of input time series needed for intelligent action. These are the channels which carry the information from the senses, as the eyes and ears, without which intelligent action would hardly be possible.

We now have all the components of the model of intelligence: a digital computer, with two types of binary input time series, and one type of binary output series. Let us arrange them in a diagram:

Model of Intelligence

	Past	Present	Future
Human Drives (as hunger, received from the body by the brain)	$H_1(1), H_1(2), H_1(3), \ldots$	$H_1(n)$,	$H_1(n+1), \ldots$
	$H_2(1), H_2(2), H_2(3), \ldots$	$H_2(n)$,	$H_2(n+1), \ldots$
	$H_3(1), H_3(2), H_3(3), \ldots$	$H_3(n)$,	$H_3(n+1), \ldots$
	----------	------	----------
	----------	------	----------
Human Senses (information flow from eyes, ears etc. to the brain)	$S_1(1), S_1(2), S_1(3), \ldots$	$S_1(n)$	$S_1(n+1), \ldots$
	$S_2(1)$, ----------	------	----------
	----------	------	----------
	----------	------	----------
	----------	------	----------
Commands to muscles (the information flow in the nerves from the brain to hand, tongue or foot)	$C_1(1), C_1(2), C_1(3), \ldots$	$C_1(n)$,	$C_1(n+1), \ldots$
	$C_2(1), \ldots$	$C_2(n)$,	$C_2(n+1), \ldots$
	----------	------	----------
	----------	------	----------
	----------	------	----------

H, S, C are all binary functions at specific intervals of time t: for instance $H_1(t=1)=1$, $H_1(t=2)=0$, $H_1(t=3)=0$, etc. The intervals $\Delta t=1$ may be 1 second, 0.1 second or the like.

[It is well known that the observing brain also acts on glands in the body to prepare it for action: the sight of food stimulates the digestive juices, the sight of fearsome things puts adrenalin into the blood, etc. Although this is not muscular action, the function, logically speaking is so similar to muscular action that it will be incorporated in the model as channels C_k, C_{k+1} etc.]

All we have to do now to finish the model of intelligence is to add a computer program that constructs, from the series $H_1, H_2 \ldots, S_1, S_2 \ldots$ and $C_1, C_2 \ldots$, the output series $C_1, C_2 \ldots$. The aim of the program is to make the future digits of H_1, H_2 equal to zero. H=0 will be assumed a satisfied drive, H=1 an active drive. It must be clear that one cannot expect a strictly causal relation between our future desires H and our present muscular activity. When one is out in the middle of a desert, it will be hard to come up with an immediate action to satisfy one's hunger or thirst. One can only ask for the best one can do, that is as many 0's as possible in the future of the H's, on the basis of one's limited capabilities.

Our program needs a theory of prediction. It is clear that such a theory should exist. People do learn to satisfy their needs by their actions. We call that skill, experience, insight, intelligence, "know how" or wisdom. One is not continuously succesful by just sheer luck. We also believe, on the basis of modern science, that the learning process takes place by information processing in the brain, analogous to what digital computers do. Does a theory of prediction exist? I will propose one[3]) to you. It is my view that such a theory, through the work on information theory, by Shannon and others, has become almost selfevident. The knowledge, that one can work with binary time series only, and arrive at a completely general theory, has certainly eliminated all mathematical difficulties. The only problem is scientific: to propose a theory that conforms to the intuitive expectations we have, as intelligent beings, about the future.

To begin, assume a binary time series f(t), and f(1)=1, f(2)=0, f(3)=1, f(4)=0, f(5)=1, f(6)=0. What is f(7), f(8), ers.? Let me write down the pattern: 101010.... One has the intuitive expectation that it will continue with 1010. Can we now find the reason for this intuitive expectation and formulate it as a general theory for all possible situations of binary time series?

Suppose one gives a specific time series to a group of scientists, experts in analysing information:
10110100101010001110.....
Let the time series be a long one, say a million digits, and one asks them to predict the future of the series. Naturally they will want to know what the time series represent, so they have some clues: is it the binary encoding of some spoken, or written language, the song of a bird, a television program? But they are told that a new born baby does not have the privilege of knowing what the world is all about. It receives signals from its eyes and ears and stomach and does not know what they mean. But in five or ten years it grows up into an intelligent being which knows a lot. Why should a group of accomplished scientists be entitled to more? So the experts start analysing the time series by all the means they have available and end up confessing that the series look very random to them, and therefore there is little to predict. One then gives them the clue that it may represent the first million digits of $\sqrt{\pi}$, written in binary digits. They try it and it fits perfectly. They will then have a strong intuitive expectation that also the future digits of the series will be represented by $\sqrt{\pi}$ Are they certain that this expectation will turn out correct? No, nothing is certain in the real world. It is simply the best prediction they have as rational beings on the basis of the available information.

This example shows the way to the general formulation of a theory of rational prediction. There would have been no reason to associate the time series with $\sqrt{\pi}$, because $\sqrt{\pi}$ would have been just one hypothesis out of a billion others one could come up with. But once a hypothesis has been tried and shown to conform to the presently given series, we have a strong intuitive confidence that this makes a valuable prediction. So the theory says: since all hypotheses are apriori equivalent, choose a handy package of hypotheses which is readily formulated, easily stored and efficiently tested. There exists one such handy package. It is: "compare the present time series with all its own pasts." For a million digit series, we have a million independent hypotheses without any effort. There is no other set of hypotheses I have been able to think of that matches this one in efficiency. I therefore hypothesize that this is the system used by all living creatures endowed with intelligence, great or small.

There are a few things to be specified before the theory is complete: The first one is that one must allow a percentage of error in the comparison of the present time series with the hypotheses. Otherwise one may find that not a single hypothesis is confirmed. The second one is that one can demand a match of the time series with the hypothesis, not over the whole series but only over an unspecified period back into the past. Then, if one has several hypotheses matching the time series, one over 10 digits one over 100 digits say, of the most recent past, one selects as the best prediction that one hypothesis which has matched the time series over the longest most recent past.

There is one more supplement to the theory which seems desirable. In our model, there is not just one time series, but a large number of series. We have presented the past, of Drives, Senses and Commands to muscles information as a two dimensional tableau of ones and zeros. To predict, we compare the present tableau not just with the tableaus we obtain by going 1,2,3 or more digits to the left. Instead, we also allow movements up and down. We investigate the match of the present tableau with all

other tableaus we obtain by going any number of steps to the left combined with any number of steps up or down.

The theory says that the brain is programmed to test n hypotheses, where the information stored by previous events in the life of the intelligent individual is n bits. It then selects the one hypothesis which conforms best to the present situation. For humans, this may require the testing of 10^9 to 10^{10} hypotheses. And don't forget, this task is performed not just once, but again and again, every time interval Δt of 0.1 seconds may be. Clearly that is an enormous computational task! There must therefore be short cuts.

To finish however, we must consider two situations. The first one is the case that the prediction leads to an increased satisfaction of the drives, an increased number of zeros. That is the easy one. The brain simply produces the corresponding commands to the muscles from its memory. This is as in the pleasant situation where one drives home after work in light traffic. One knows what to do. The second possibility is that the prediction is one of increased anxiety. "One enjoys one's dinner at home and notices through the window one's creditors converge on the front door" Or: "one enters one's home and finds a boa constrictor in the living room" What is now the model for the automatic instruction to the muscles? These questions are not answered as readily, and clearly require more discussion.

April 24, 1981

References

1 C. E. Shannon and W. Weaver "The Mathematical Theory of Communication," The University of Illinois Press, Urbana 1949

2 W. MacDougall "An Introduction of Social Psychology," Barnes and Noble, New York 1960 (originally 1908)

3 P. J. vanHeerden "The Foundation of Empirical Knowledge," Wistik, Wassenaar, Netherlands, 1968