

# Communication Amongst Automata (1962)

Heinz von Foerster



Biological Computer Lab Publication #81

BCL Fiche #90, #22–29

For those of us who are actively engaged in research concerned with systems of high complexity and who think about their implementation and their future application, it is quite obvious that today we are in the midst of an era which provides the ideal conditions for the fast evolution of the automaton with mind-like behavior. Thus, I appreciate very much the occasion to be permitted to give you, the psychiatrist, a short report about these developments, because — I believe — in not too distant a future it will be the psychiatrists who will be confronted with problems arising from the interaction of man with his new baby, the “intelligent” automaton.

Furthermore, I hope that the discussion of these new machines and their potentialities will give me an appropriate vehicle to present some of the fundamental concepts in an interaction process commonly referred to as “communication” concepts which, I believe, will hold for any communication process, whether it takes place between machines, beasts and man, or between all their possible combinations.

My first task in this presentation is a rehabilitation of the term “automaton.” Unfortunately, in formal discussion, but also in recent literature, journal articles and in the press, you will find the terms “automata” and “robots” freely interchanged as if they would refer to one and the same thing. This, however, is not the case. While “robot” is derived from the Czech word *robotnic* (worker), became popular through Capek’s delightful play *Rossum’s Universal Robots*, and refers to a stupid mechanism carrying out without its own initiative all that it is commanded to do, “automaton” is derived from the

Greek *automatizein* (to act according to one’s own will), and thus refers to a gadget on a much higher level of sophistication. Indeed, if you care to look up “automaton” in a dictionary<sup>1</sup> you will find that an automaton is “... a contrivance constructed to act *as if* spontaneously, through concealed motive power.” It may be argued that this definition describes still a pedestrian gadget, because with patience and skill we may “reveal” the concealed mechanism. However, the situation changes drastically, if — for some reason or another — we are in principle unable to reveal that hidden mechanism. Under these circumstances we are forced to drop the “as if” in the above definition and we have a truly “free” system before us which acts on its “own will.” It may, perhaps, amuse you to note that Aristotle used the term “automaton” in the latter sense.<sup>2</sup> I presume that a bad translation in the early nineteenth century of the famous passage in “De Motu” shifted its meaning to its weaker usage of today.

At this point you may rightly ask how such systems can ever be built. Unfortunately, a tight proof of my assertion of the feasibility of such systems would take up a one-semester seminar, thus, I hope you will believe me if I can assure you that such systems can be built, if they are made up of elementary components which fulfill the following four conditions:<sup>3</sup>

1. The elementary component must be energetically autonomous. That is, it should not receive its energy via the information channels from other such units, but should be able to extract energy from its environment.
2. The rules by which the elementary component

<sup>1</sup>American Collegiate Dictionary. New York: Harper & Brothers, 1948. (It may be noted that Webster’s New World Dictionary, 1959, does not use the “as if.”)

<sup>2</sup>Aristotle: *De Motu Animalium*. In Smith, J. A. and Ross, W. D. (Trans.): *The Works of Aristotle*, Vol. V, 701b–703a. Oxford, 1958.

<sup>3</sup>Pask, G., and Von Foerster, H.: *Cybernetica*, 3: 258, 1960. *Cybernetics*, 4: 20, 1961.

handles the information presented to it at a particular time must depend on earlier states of the component and also on the frequency of its use.

3. The elementary component must be able to make trials. That is, it should generate stimuli to other units (or the system's "environment"), these stimuli being not necessarily responses to stimuli of the elementary component.
4. If conditions 1 and 3 are fulfilled, one can finally demand that the threshold for trial making be lowered, if the environment of the elementary component becomes energetically depleted.

By going carefully through this list of properties you may have spotted three important features. Number one, that these properties may well be attributed to a cortical neuron, if only some of its outstanding functional properties are taken into consideration. This should not come as a surprise, since we know that most of these two-legged admirable automata — in the Aristotelian sense — are equipped with about  $10^{10}$  such elementary components. Number two, that condition 4 stipulates — if not explicitly — a certain "personal" or "microscopic" [sic] goal which the elementary component "seeks" to attain by its trial making activity. This goal is, of course, the maintenance of an energetically resourceful environment, in spite of the component's metabolic activity. As we shall see in a moment, this paradox is resolved by the component's ability to communicate with other elements in its environment. And, finally, number three, that I can be accused of building an automaton by using as elementary components automata. I am not going to refuse this argument: on the contrary, I wholeheartedly agree, with one reservation however, namely, that I have given the necessary and sufficient prescriptions to construct such elementary components. Indeed, there are many versions of electronic realizations of such components in existence today, from sizes of a couple of cubic inches down to about 2 cubic millimeters, and in costs ranging from \$5000 for very sophisticated devices down to a couple of cents apiece, barely fulfilling the points mentioned above. It is, however, not the particular component which is worth mentioning here. A single component in itself has no value whatsoever. Only due to the fact that they are capable of communicating with each other are they in a position to form coalitions by which these elements can achieve jointly what all elements separately would never be able to accomplish.

The secret behind the advantage for the individual to join a coalition lies, of course, in a super additive composition rule applicable for communicating elements. By

this I refer to the old — but unfortunately inaccurate — saying that "the whole is more than the sum of its parts." Although this statement has been under heavy attack by positivists, operationalists, *etc.*, if put properly it emerges as a most important guiding principle in the theory and technology of self organizing and adaptive systems. Properly formulated we would say today that to a set of communicating elements we have to apply a super additive composition rule, because "a *measure* of the sum of its parts is larger than the sum of the *measure* of its parts." Consider the function  $\phi$  as a measure; then for the two parts  $x$  and  $y$  we have:

$$\phi(x+y) > \phi(x) + \phi(y)$$

This equation is nothing else but my definition of a coalition being put into precise, mathematical language. If you have any doubt as to the existence of such a measure which will satisfy the above equation, I suggest, *e.g.*, using for  $\phi = ()^2$ , that is "taking the square." We have

$$(x+y)^2 > x^2 + y^2$$

which is obviously true, because  $(x+y)^2 = x^2 + y^2 + 2xy$ , hence the left hand side of the inequality always exceeds the right hand side by an amount of  $2xy$ . Perhaps the following biological example will make my point of a super-additive composition rule even clearer:

$$\Phi(\text{👤}) > \Phi(\text{👋}) + \Phi(\text{👤}) .$$

The most important example of such a measure in connection with my topic is one which has been developed in information theory.<sup>4</sup> It is the concept of "certainty" or, as it is often referred to, as "neg-entropy," symbolized by  $-H$ . One of the most important findings in this theory is that the certainty of a joint event  $-H(x+y)$  is always larger or equal to the sum of the certainties of the individual events,  $-H(x) - H(y)$ , equality being the case only for completely independent events. Or expressed the other way around:

$$H(x+y) < H(x) + H(y)$$

The uncertainty of a joint event is always smaller or equal to the sum of the uncertainties of the individual events. Let me illustrate this on an oversimplified example, which however, can be developed into a calculus of general validity.

Assume that there is a highly specialized physicist  $P$ , who knows only one proposition:

$$x = \text{"electrons are negative"}$$

<sup>4</sup>Shannon, C. E. and Weaver, W.: *The Mathematical Theory of Communication*. Urbana: University of Illinois Press, 1948. p. 21.

Assume furthermore that there is a highly specialized biologist  $B$ , who also knows only one proposition:

$$y = \text{"elephants are gray"}$$

Using the conventional logical symbols

$$\vee \quad \& \quad -$$

for

$$\text{"or"} \quad \text{"and"} \quad \text{"negation"}$$

respectively, the physicist's knowledge of the universe can be stated as follows:

$$x \& (y \vee \bar{y})$$

which, in words, says: "electrons are negative; and elephants are gray or elephants are not gray."

While the biologist's picture of the universe is:

$$y \& (x \vee \bar{x})$$

which, in words, says: "elephants are gray; and electrons are negative or electrons are not negative."

This situation of the knowledge of the two independent scholars can be neatly expressed in form of a "truth-table" associating the numbers 1 and 0 with "true" and "false" respectively for the propositions  $x$  and  $y$  and the associated logical functions as expressed above. The truth-table for the two gentlemen reads thus as follows:

|     |     | $P$                     | $B$                     |
|-----|-----|-------------------------|-------------------------|
| $x$ | $y$ | $x \& (y \vee \bar{y})$ | $y \& (x \vee \bar{x})$ |
| 0   | 0   | 0                       | 0                       |
| 0   | 1   | 0                       | 1                       |
| 1   | 0   | 1                       | 0                       |
| 1   | 1   | 1                       | 1                       |

In other words, the physicist will always believe to have made a true statement when  $x$  is true, independent of whether  $y$  is true or false. Similarly the biologist, *mutatis mutandis*.

However, if the two gentlemen are forming a coalition by establishing, say, a "Biophysical Society" the truth-table of the Society is clearly dictated by the knowledge of both scholars together and thus reads:

| Bph. S. |     |          |
|---------|-----|----------|
| $x$     | $y$ | $x \& y$ |
| 0       | 0   | 0        |
| 0       | 1   | 0        |
| 1       | 0   | 0        |
| 1       | 1   | 1        |

Comparing the truth-table of the society with the truth-tables of the individuals one easily sees that the number of instances in which the response "true" is elicited for a particular state of the universe has decreased after coalition, hence the society is less credulous than the individuals, its uncertainty is diminished and it will respond with "true" only if the universe is adequately described: "electrons are negative and elephants are gray." These considerations can be expanded considerably and it is not difficult to show that with a sufficient number of elements each of which possesses only a very limited knowledge, an arbitrary degree of certainty with respect to their universe can be obtained if these elements are capable of exchanging the little bit of knowledge they possess, or, in other words, if they form a coalition by communication.

Amongst the flood of examples which could be cited in support of this thesis, let me briefly mention only the strikingly increased survival value for living organisms when associated in coalitions. The number on unicellular organisms on this planet is about of the order of  $10^{17}$ . This is quite an impressive number if one considers that the age of the universe expressed in seconds.

Although there are by far less insects on this globe than unicellular organisms, the number of cells which have organized themselves into insects is of the order of  $10^{20}$ . Hence, a cell participating in a "coalition" called, e.g., "mosquito," is about a thousand times more stable than being isolated. However, these numbers are dwarfed, if we look at a cellular aggregate of the size of *Homo sapiens*. With each of us representing a colony of approximately  $3 \cdot 10^{15}$  cells, the participants of this meeting comprise more cells than all unicellular organisms on this globe, and with  $10^{25}$  cells in "human coalitions" mankind represent probably more cells than the rest of all living organisms.

Up to this point I have only discussed the necessity for information flow in autonomous, decision making systems. I hope that I have made sufficiently plausible that a system composed of communicating automata provides each automaton with a higher payoff function — e.g., survival value — than would be possible in a mere set of automata, and also that a system of automata closely linked to each other by active communication channels — a coalition — can again be considered as a single automaton of higher complexity. However, I have not mentioned with a single word that which is communicated amongst these automata or amongst their elementary components. Indeed, what are they talking about?

It is impossible to answer this question if an automaton is considered to be an isolated entity. In order that

this question makes any sense at all, we have to immerse the automaton into an environment with many possible states, or — to be more poetic — where the wind blows, the sun shines, rocks tumble, water splashes; in other words, where something is going on. In addition, we may allow this environment to contain other automata, either of the same kind or of different make-up. In order that these automata show some stability in this “hostile” environment, it is clear that they have to discover some order in this environment. In an absolute chaos their survival is questionable. When I use the word “order” I simply mean that in this environment not everything happens that could happen. In our environment, for instance, we find that most things maintain their shapes, fall downwards and do not move in zig-zag motions through space, *etc.* In other words, the transition probabilities for certain state sequences are very close to unity, while others are vanishingly small. This is just another way of saying that there are “Laws of Nature” and our textbooks of physics, chemistry, astronomy and so, are nothing else but a codification of these laws.

After these preliminaries on the structure of the environment it is now obvious that in order for our automata to survive, they have to crack the code of their environment’s intrinsic order. This they have accomplished if they can find a solution for an internal representation of the order of their surrounding universe. Although it is to a certain degree irrelevant in which code this representation is accomplished — sequences of electric pulses, sound frequencies, black marks on a white background, wiggly grooves on a black disk, changing magnetic patterns on a flexible tape, *etc.* — it is of great importance that this code is shared by many elements comprising the automaton, because — as we have seen earlier — it is the joint knowledge of the elements of the system which makes the system wiser than the sum of the wisdom of its parts. This answers the question as to what is communicated: it is information about the structure of the environment of the automaton.

With these remarks I could conclude my discussion on communication among automata,. However, it may be profitable to illustrate the principles presented herein with a brief allegory. In Figure 1 we have an automaton which lives on gasoline which he consumes when it is fed to him in small cans. These he can measure and weigh. If they are too large for his consumption, or if there are too light, *i.e.*, empty, he kicks them over with his leg. Furthermore, he has four lamps arranged in a square which he can turn on and off, one or more at a time. With these he can communicate his needs. Clearly, with any of his four lamps in two possible states — on or off — he has precisely  $2^4 = 16$  different “words” in his vocabulary.

In the present state of affairs, whoever, these words have as yet no “meaning” whatsoever. This is precisely what we are going to teach him. To this end we have to invent some environmental rules. Let us agree on the following convention (Figure 2): the coordinates of his square lamp box are to represent upwards = size, and horizontal = weight.

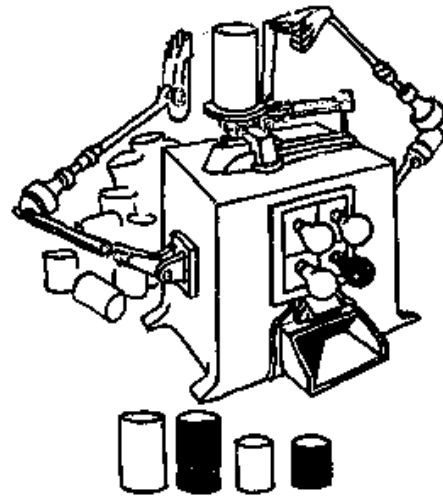


FIGURE 1

In other words, if he lights up, say, the upper right lamp, we feed him a tall, filled can, which he will reject, of course. Such a trial will be 100% unsuccessful. If he lights up two lamps simultaneously, say the lower left and right lamps, we will feed him with equal probability either a small empty can or a small filled can. This would make him successful 50% of the time. Clearly, what he has to learn is to light up the lower right lamp only, which causes his environment to feed him the desired small, filled can.

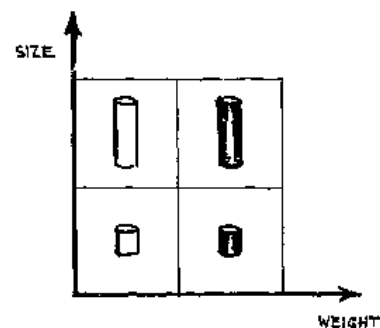


FIGURE 2

Since our automaton is constructed out of elementary components which follow the points 1 to 4 mentioned earlier, he will, after a series of more or less successful trials, resume the habit of turning on only the lower right lamp when hungry. He knows now the “meaning” of this

word; it means “digestible food.” At this stage it may become monotonous to feed this critter whenever he turns on his lamp. We do not gain anything by watching this device which has turned into a boring, deterministic system.

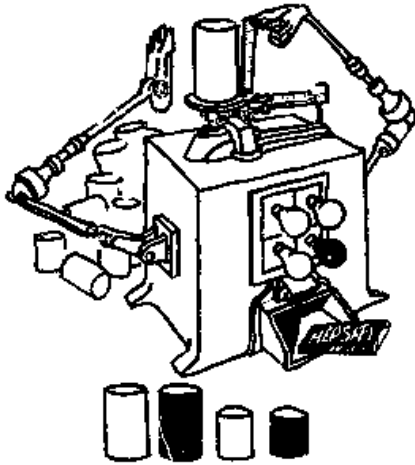


FIGURE 3

However, at that instant the automaton presents us with his waste products which happens to be a chocolate bar (Figure 3). This, of course, changes the situation for us, because it offers certain advantages for us if we are going to form a coalition with this system. Again this may go on for a while — but who wants to live on chocolate forever? Thus we may replace ourselves by an automaton that can't do anything with gasoline, but who thrives on chocolate bars (Figure 4). Furthermore, this new fellow has eyes to recognize light-signals and to distinguish gasoline cans. Wheels permit him to move freely in his

environment which consists again of large and small, empty and filled, gasoline cans. He, too, will learn to understand the language of our primary automaton, and he will too, after some time of adaptation, appreciate the advantages of a coalition. Before long you will see these two automata joined together (Figure 5), the one acting as a stomach, cracking up the raw-products and transforming them into digestible foodstuff for his permanent partner, who acts as sensor and effector.

May I assure you that there exist today neither conceptual nor technological difficulties to realize such automata in mechanical and electronic hardware. Indeed, we have the theoretical and technological know-how to construct systems in comparison to which the two characters of my little allegory would look like simpletons. Given a bit more time, I venture to say that in comparison with these future automata, even we may look like simpletons.

Since man is limited his capacity to process information and to make complex decisions, and since man's environment becomes more and more intricate, because it is more and more defined by man's own complexity, it is not absurd to predict that within one generation adaptive, decision-making automata will play a decisive role in charting the course of human events.

I am sure that most of us agree that we are today in the midst of a major transformation of the human condition, and I am convinced that in this period of transition the role of the psychiatrist will be of paramount importance, because it will be he who will have to deal with the frustrations resulting from our incapacity to communicate with these future automata.

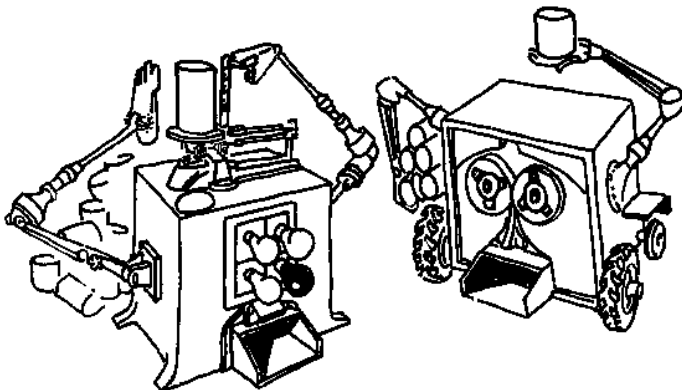


FIGURE 4

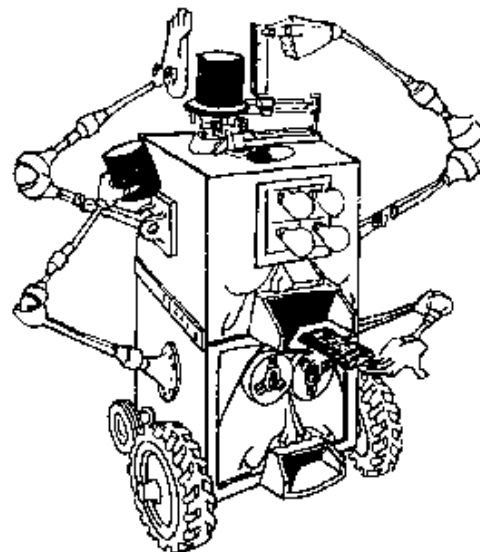


FIGURE 5