Theory of
Self-Reproducing Automata

JOHN VON NEUMANN

edited and completed by Arthur W. Burks

University of Illinois Press
URBANA AND LONDON 1966

Fourth Lecture

THE ROLE OF HIGH AND OF EXTREMELY HIGH COMPLICATION

Comparisons between computing machines and the nervous systems. Estimates of size for computing machines, present and near future.

Estimates for size for the human central nervous system. Excursus about the "mixed" character of living organisms. Analog and digital elements. Observations about the "mixed" character of all componentry, artificial as well as natural. Interpretation of the position to be taken with respect to these.

Evaluation of the discrepancy in size between artificial and natural automata. Interpretation of this discrepancy in terms of physical factors. Nature of the materials used.

The probability of the presence of other intellectual factors. The role of complication and the theoretical penetration that it requires.

Questions of reliability and errors reconsidered. Probability of individual errors and length of procedure. Typical lengths of procedure for computing machines and for living organisms—that is, for artificial and for natural automata. Upper limits on acceptable probability of error in individual operations. Compensation by checking and self-correcting features.

Differences of principle in the way in which errors are dealt with in artificial and in natural automata. The "single error" principle in artificial automata. Crudeness of our approach in this case, due to the lack of adequate theory. More sophisticated treatment of this problem in natural automata: The role of the autonomy of parts. Connections between this autonomy and evolution.

After the broad general discussions of the last two lectures I would like to return to the subject of the specific automata which we know. I would like to compare artificial automata, specifically computing machines, with natural automata, particularly the human nervous system. In order to do this, I must say a few things in both cases about components and I must make certain comparisons of sizes.

As I mentioned before, in estimating the size of the human nervous system one is limited to a figure which is not very well established, but which is probably right in its order of magnitude. This is the

statement that there are 10^{10} neurons in the human brain. The number of nerves present elsewhere in the human organism is probably much smaller than this. Also, a large number of these other nerves originate in the brain anyway. The largest aggregation of nerves of the periphery is on the retina, and the optic nerve going from the retina to the brain is part of the brain.

Compared to this, the number of vacuum tubes involved in the computing machines we know of is very small, a million times smaller. The largest existing computing machine, the ENIAC, has 2×10^4 vacuum tubes. Another large computing machine, the SSEC, which belongs to the IBM Company, contains a mixture of vacuum tubes and relays, about 10 thousand of each. The fastest computing machines now under construction are designed to have several thousand vacuum tubes, perhaps 3 thousand. The reason for this difference in size between the ENIAC and the fast machines now under construction is a difference in the treatment of memory, which I will discuss later.

So the human nervous system is roughly a million times more complicated than these large computing machines. The increase in complexity from these computing machines to the central nervous system is more than the increase in complexity from a single vacuum tube to these computing machines. Even measuring complexity on a logarithmic scale, which is highly generous, we have not yet come half the way. I think that in any sensible definition of complexity, it would be much less than half way.

There is, however, a factor in favor of these machines: they're faster than the human brain. The time in which a human nerve can respond is about $\frac{1}{2}$ millisecond. However, that time is not a fair measure of the speed of the neuron, because what matters is not the time in which the neuron responds, but the time in which it recovers, the time from one response to the next potential response. That time is, at best, 5 milliseconds. In the case of a vacuum tube it's difficult to estimate the speed, but present designs call for repetition rates which are not much in excess of a million per second.

Thus the nervous system has a million times as many components as these machines have, but each component of the machine is about 5 thousand times faster than a neuron. Counting what can be done, hour by hour, the nervous system outperforms the machine by a factor of roughly 200. This estimate, however, favors the automaton, because an n-fold increase in size brings much more than an n-fold increase in what can be done. What can be done is a matter of the interrelationships between the components, and the number of

interrelationships increases with the square of the number of components. And apart from this, what can be done depends on certain minima. Below a certain minimum level of complexity you cannot do a certain thing, but above this minimum level of complexity you can do it.

[Von Neumann next compared the human nervous system and computers with respect to volume. The decisive factor is the space in which the control and amplifying functions are performed. In the case of the vacuum tube this is essentially the space between the cathode and the control grid, which is of the order of magnitude of a millimeter. In the case of the nerve cell it is the thickness of the nerve membrane, which is of the order of 1 micron. The ratio in size is about 1000 to 1, and this is also the ratio in voltage, so that the intensity of the field which is used for control and amplification is about the same in the vacuum tube and the nerve cell. This means that differences in total energy dissipation are mainly due to differences in size. "A discrepancy of 10^3 in linear size means a discrepancy of 10^9 in volume, and probably a not very different discrepancy in energy." See also *Collected Works* 5.299–302 and *The Computer and the Brain* 44–52.

He then calculated the energy which is dissipated "per elementary act of information, that is, per elementary decision of a two-way alternative and per elementary transmittal of 1 unit of information." He did this for three cases: the thermodynamical minimum, the vacuum tube, and the neuron.

In the third lecture he said that thermodynamical information is measured by the logarithm, to the base two, of the number of alternatives involved. The thermodynamical information in the case of two alternatives is thus one, "except that this is not the unit in which you measure energy. Entropy is energy only if you specify the temperature. So, running at low temperature you can say what energy should be dissipated." He then computed the thermodynamical minimum of energy per elementary act of information from the formula $kT \log_e N$ ergs, where k is Boltzmann's constant (1.4×10^{-16} ergs per degree), T is the temperature in absolute units, and N is the number of alternatives. For a binary act $N = 2$, and taking the temperature to be about 300 degrees absolute, he obtained 3×10^{-14} ergs for the thermodynamical minimum.

Von Neumann then estimated that the brain dissipates 25 watts, has 10^{10} neurons, and that on the average a neuron is activated about 10 times per second. Hence the energy dissipation per binary act in a nerve cell is roughly 3×10^{-3} ergs. He estimated that a vacuum tube dissipates 6 watts, is activated about 100,000 times per second, and thus dissipates 6×10^2 ergs per binary act.]

So our present machinery is about 200 thousand times less efficient than the nervous system is. Computing machines will be improved in the next few years, perhaps by replacing vacuum tubes with amplifying crystals, but even then they will be of the order of 10 thousand times less efficient than the nervous system. The remarkable thing, however, is the enormous gap between the thermodynamical minimum (3×10^{-14} ergs) and the energy dissipation per binary act in the neuron (3×10^{-3} ergs). The factor here is 10^{11}. This shows that the thermodynamical analysis is missing a large part of the story. Measured on a logarithmic scale, the gap between our instrumentation, which is obviously amateurish, and the procedures of nature, which show a professional touch, is about half the gap between the best devices we know about and the thermodynamical minimum. What this gap is due to I don't know. I suspect that it's due to something like a desire for reliability of operation.

Thus, for an elementary act of information, nature does not use what, from the point of view of physics, is an elementary system with two stable states, such as a hydrogen atom. All the switching organs used are much larger. If nature really operated with these elementary systems, switching organs would have dimensions of the order of a few angstroms, while the smallest switching organs we know have dimensions of the order of thousands or tens of thousands of angstroms. There is obviously something which forces one to use organs several orders of magnitude larger than is required by the strict thermodynamical argument. Thus, though the observation that information is entropy tells an important part of the story, it by no means tells the whole story. There is a factor of 10^{11} still to be accounted for.

[Von Neumann then discussed memory components. Vacuum tubes, which are switching organs, may be used for memory. But since the standard circuit for storing a binary digit has two tubes, and additional tubes are needed for transmitting the information in and out, it is not feasible to build a large memory out of vacuum tubes. "The actual devices which are used are of such a nature that the store is effected, not in a macroscopic object like a vacuum tube, but in something which is microscopic and has only a virtual existence." Von Neumann describes two devices of this sort: acoustic delay line storage and cathode ray tube storage.

An acoustic delay line is a tube which is filled with a medium such as mercury and which has a piezo-electric crystal at each end. When the transmitting crystal is stimulated electrically, it produces an acoustic wave that travels through the mercury and causes the receiving crystal to produce an electrical signal. This signal is amplified, reshaped, and retimed and sent to the transmitting crystal again.

This acoustic-electrical cycle can be repeated indefinitely, thereby providing storage. A binary digit is represented by the presence or absence of a pulse at a given position at a given time, and since the pulses circulate around the system, the digit is not stored in any fixed position. "The thing which remembers is nowhere in particular."

Information may be stored in a cathode ray tube in the form of electric charges on the inside surface of the tube. A binary digit is represented by the charge stored in a small area. These charges are deposited and sensed by means of the electron beam of the cathode ray tube. Since the area associated with a given binary digit must be recharged frequently, and since this area may be moved by changing the position of the electron beam, this memory is also virtual. "The site of the memory is really nowhere organically, and the mode of control produces the memory organ in a virtual sense, because no permanent physical changes ever occur."]

There's therefore no reason to believe that the memory of the central nervous system is in the switching organs (the neurons). The size of the human memory must be very great, much greater than 10^{10} binary units. If you count the impressions which a human gets in his life or other things which appear to be critical, you obtain numbers like 10^{15}. One cannot place much faith in these estimates, but I think it likely that the memory capacity of the human nervous system is greater than 10^{10}. I don't know how legitimate it is to transfer our experience with computing machines to natural systems, but if our experience is worth anything it is highly unlikely that the natural memory should be in switching organs or should consist of anything as unsophisticated and crude as the modification of a switching organ. It has been suggested that memory consists in a change of threshold at a synapse. I don't know if this is true, but the memory of computing machines does not consist of bending a grid. A comparison between artificial automata and the central nervous system makes it probable that the memory of the latter is more sophisticated and more virtual than this. Therefore, I think that all guesses about what the memory of the human organism is, and where it sits, are premature.

Another thing of which I would like to talk is this. I have been talking as if a nerve cell were really a pure switching organ. It has been pointed out by many experts in neurology and adjacent fields that the nerve cell is not a pure switching organ but a very delicate continuous organ. In the lingo of computing machinery one would say it is an analog device that can do vastly more than transmit or not transmit a pulse. There is a possible answer to this, namely, that vacuum tubes, electromechanical relays, etc. are not switching devices

either, since they have continuous properties. They are all characterized by this, however, that there is at least one way to run them where they have essentially an all-or-none response. What matters is how the component runs when the organism is functioning normally. Now nerve cells do not usually run as all-or-none organs. For instance, the method of translating a stimulus intensity into a frequency of response depends on fatigue and the time of recovery, which is a continuous or analog response. However, it is quite clear that the all-or-none character of a neuron is a very important part of the story.

The human organism is not a digital organ either, though one part of it, the nervous system, is essentially digital. Almost all the nervous stimuli end in organs which are not digital, such as a contracting muscle or an organ which causes secretions to produce a chemical. To control the production of a chemical and rely on the diffusion rate of a chemical is to employ a much more sophisticated analog procedure than we ever use in analog computing machines. The most important loops in the human system are of this nature. A system of nervous stimuli goes through a complicated network of nerves and then controls the operation of what is essentially a chemical factory. The chemicals are distributed by a very complicated hydrodynamical system, which is completely analog. These chemicals produce nervous stimuli which travel in a digital manner through the nervous system. There are loops where this change from digital into analog occurs several times. So the human organism is essentially a mixed system. But this does not decrease the necessity for understanding the digital part of it.

Computing machines aren't purely digital either. The way we run them now, their inputs and outputs are digital. But it's quite clear that we need certain non-digital inputs and outputs. It's frequently desirable to display the result, not in digits, but, say, as a curve on an oscilloscope screen. This is an analog output. Moreover, I think that the important applications of these devices will come when you can use them to control complicated machinery, for example, the flight of a missile or of a plane. In this case the inputs will come from an analog source and the outputs will control an analog process. This whole trans-continuous alternation between digital and analog mechanisms is probably characteristic of every field.

The digital aspect of automata should be emphasized at the present time, for we now have some logical tools to deal with digital mechanisms, and our understanding of digital mechanisms is behind our understanding of analog mechanisms. Also, it appears that digital

mechanisms are necessary for complicated functions. Pure analog mechanisms are usually not suited for very complicated situations. The only way to handle a complicated situation with analog mechanisms is to break it up into parts and deal with the parts separately and alternately, and this is a digital trick.

Let me now come to the following question. Our artificial automata are much smaller than natural automata in what they do and in the number of components they have, and they're phenomenally more expensive in terms of space and energy. Why is this so? It's manifestly hopeless to produce a true answer at the present time: We can hardly explain why two objects are different if we understand one a little and the other not at all. However, there are some obvious discrepancies in the tools with which we operate, which make it clear that we would have difficulty in going much further with these tools.

The materials which we are using are by their very nature not well suited for the small dimensions nature uses. Our combinations of metals, insulators, and vacuums are much more unstable than the materials used by nature; that they have higher tensile strengths is completely incidental. If a membrane is damaged it will reconstruct itself, but if a vacuum tube develops a short between its grid and cathode it will not reconstruct itself. Thus the natural materials have some sort of mechanical stability and are well balanced with respect to mechanical properties, electrical properties, and reliability requirements. Our artificial systems are patchworks in which we achieve desirable electrical traits at the price of mechanically unsound things. We use techniques which are excellent for fitting metal to metal but are not very good for fitting metal to vacuum. To obtain millimeter spacings in an inaccessible vacuum space is a great mechanical achievement, and we will not be able to decrease the size by large factors here. And so the differences in size between artificial and natural automata are probably connected essentially with quite radical differences in materials.

[Von Neumann proceeded to discuss what he thought was a deeper cause of the discrepancy in size between natural and artificial automata. This is that many of the components of the natural system serve to make the system reliable. As he noted in the third lecture, actual computing elements function correctly with a certain probability only, not with certainty. In small systems the probability that the whole system will behave incorrectly is relatively small and may often be neglected, but this is not the case with large systems. Thus error considerations become more important as the system becomes more complex.

Von Neumann made some very rough calculations to justify this

conclusion. Assuming that the system is designed in such a way that the failure of a single element would result in failure of the whole system, he calculated the error probability required for a given mean free path between system errors. For the human nervous system he used the following figures: 10^{10} neurons; each neuron activated 10 times per second on the average; a mean free path between fatal errors of 60 years (the average life span). Since 60 years is about 2×10^9 seconds, the product of these numbers is 2×10^{20}. Hence an error probability of 0.5×10^{-20} for each activation of an element is required under these assumptions. For a digital computer he used the figures: 5×10^3 vacuum tubes, 10^5 activations per tube per second, and a desired mean free path between system errors of 7 hours (about 2×10^4 seconds). An error probability of 10^{-13} per tube activation is required for this degree of reliability. Compare the calculations at *Collected Works* 5.366–367.

He pointed out that vacuum tubes, and artificial components generally, do not have an error probability as low as 10^{-13}, and that neurons probably do not either. We try to design computing machines so that they will stop when they make an error and the operator can then locate it and correct it. For example, a computer may perform a certain operation twice, compare the results, and stop if the results differ.]

It's very likely that on the basis of the philosophy that every error has to be caught, explained, and corrected, a system of the complexity of the living organism would not run for a millisecond. Such a system is so well integrated that it can operate across errors. An error in it does not in general indicate a degenerative tendency. The system is sufficiently flexible and well organized that as soon as an error shows up in any part of it, the system automatically senses whether this error matters or not. If it doesn't matter, the system continues to operate without paying any attention to it. If the error seems to the system to be important, the system blocks that region out, by-passes it, and proceeds along other channels. The system then analyzes the region separately at leisure and corrects what goes on there, and if correction is impossible the system just blocks the region off and by-passes it forever. The duration of operability of the automaton is determined by the time it takes until so many incurable errors have occurred, so many alterations and permanent by-passes have been made, that finally the operability is really impaired. This is a completely different philosophy from the philosophy which proclaims that the end of the world is at hand as soon as the first error has occurred.

To apply the philosophy underlying natural automata to artificial

automata we must understand complicated mechanisms better than we do, we must have more elaborate statistics about what goes wrong, and we must have much more perfect statistical information about the milieu in which a mechanism lives than we now have. An automaton can not be separated from the milieu to which it responds. By that I mean that it's meaningless to say that an automaton is good or bad, fast or slow, reliable or unreliable, without telling in what milieu it operates. The characteristics of a human for survival are well defined on the surface of the earth in its present state, though for most types of humans you must actually specialize the situation a little further than this. But it is meaningless to argue how the human would survive on the bottom of the ocean or in a temperature of 1000 degrees centigrade. Similarly, in discussing a computing machine it is meaningless to ask how fast or how slow it is, unless you specify what type of problems will be given to it.

It makes an enormous difference whether a computing machine is designed, say, for more or less typical problems of mathematical analysis, or for number theory, or combinatorics, or for translating a text. We have an approximate idea of how to design a machine to handle the typical general problems of mathematical analysis. I doubt that we will produce a machine which is very good for number theory except on the basis of our present knowledge of the statistical properties of number theory. I think we have very little idea as to how to design good machines for combinatorics and translation.

What matters is that the statistical properties of problems of mathematical analysis are reasonably well known, and as far as we know, reasonably homogeneous. Consider some problems in mathematical analysis which look fairly different from each other and which by mathematical standards are very different: finding the roots of an equation of the tenth order, inverting a matrix of the twentieth order, solving a proper value problem, solving an integral equation, or solving an integral differential equation. These problems are surprisingly homogeneous with respect to the statistical properties which matter for a computing machine: the fraction of multiplications to other operations, the number of memory references per multiplication, and the optimal hierarchic structure of the memory with respect to access time. There's vastly less homogeneity in number theory. There are viewpoints under which number theory is homogeneous, but we don't know them.

So, it is true for all these automata that you can only assign them a value in combination with the milieu which they have to face. Natural automata are much better suited to their milieu than any

artifacts we know. It is therefore quite possible that we are not too far from the limits of complication which can be achieved in artificial automata without really fundamental insights into a theory of information, although one should be very careful with such statements because they can sound awfully ridiculous 5 years later.

[Von Neumann then explained why computing machines are designed to stop when a single error occurs. The fault must be located and corrected by the engineer, and it is very difficult for him to localize a fault if there are several of them. If there is only one fault he can often divide the machine into two parts and determine which part made the error. This process can be repeated until he isolates the fault. This general method becomes much more complicated if there are two or three faults, and breaks down when there are many faults.]

The fact that natural organisms have such a radically different attitude about errors and behave so differently when an error occurs is probably connected with some other traits of natural organisms, which are entirely absent from our automata. The ability of a natural organism to survive in spite of a high incidence of error (which our artificial automata are incapable of) probably requires a very high flexibility and ability of the automaton to watch itself and reorganize itself. And this probably requires a very considerable autonomy of parts. There is a high autonomy of parts in the human nervous system. This autonomy of parts of a system has an effect which is observable in the human nervous system but not in artificial automata. When parts are autonomous and able to reorganize themselves, when there are several organs each capable of taking control in an emergency, an antagonistic relation can develop between the parts so that they are no longer friendly and cooperative. It is quite likely that all these phenomena are connected.

Fifth Lecture

RE-EVALUATION OF THE PROBLEMS OF COMPLICATED AUTOMATA— PROBLEMS OF HIERARCHY AND EVOLUTION

Analysis of componentry and analysis of integration. Although these parts have to appear together in a complete theory, the present state of our information does not justify this yet.

The first problem: Reasons for not going into it in detail here. Questions of principle regarding the nature of relay organs.

The second problem: Coincides with a theory of information and of automata. Reconsideration of the broader program regarding a theoretical discussion of automata as indicated at the end of the second lecture.

Synthesis of automata. Automata which can effect such syntheses.

The intuitive concept of "complication." Surmise of its degenerative character: In connection with descriptions of processes by automata and in connection with syntheses of automata by automata.

Qualifications and difficulties regarding this concept of degeneracy.

Rigorous discussion: Automata and their "elementary" parts. Definition and listing of elementary parts. Synthesis of automata by automata. The problem of self-reproduction.

Main types of constructive automata which are relevant in this connection: The concept of a general instruction. The general constructive automaton which can follow an instruction. The general copying automaton. The self-reproducing combination.

Self-reproduction combined with synthesis of other automata: The enzymatic function. Comparison with the known major traits of genetic and mutation mechanisms.

The questions on which I've talked so far all bear on automata whose operations are not directed at themselves, so that they produce results which are of a completely different character than themselves. This is obvious in each of the three cases I have referred to.

It is evident in the case of a Turing automaton, which is a box with a finite number of states. Its outputs are modifications of another entity, which, for the sake of convenience, I call a punched tape.

This tape is not itself an object which has states between which it can move of its own accord. Furthermore, it is not finite, but is assumed to be infinite in both directions. Thus this tape is qualitatively completely different from the automaton which does the punching, and so the automaton is working into a qualitatively different medium.

This is equally true for the automata discussed by McCulloch and Pitts, which are made of units, called neurons, that produce pulses. The inputs and outputs of these automata are not the neurons but the pulses. It is true that these pulses may go to peripheral organs, thereby producing entirely different reactions. But even there one primarily thinks, say, of feeding the pulses into motor or secretory organs, so it is still true that the inputs and outputs are completely different from the automaton itself.

Finally, it is entirely true for computing machines, which can be thought of as machines which are fed, and emit, some medium like punched tape. Of course, I do not consider it essentially different whether the medium is a punched card, a magnetic wire, a magnetized metal tape with many channels on it, or a piece of film with points photographed on it. In all these cases the medium which is fed to the automaton and which is produced by the automaton is completely different from the automaton. In fact, the automaton doesn't produce any medium at all; it merely modifies a medium which is completely different from it. One can also imagine a computing machine with an output of pulses which are fed to control completely different entities. But again, the automaton is completely different from the electrical pulses it emits. So there's this qualitative difference.

A complete discussion of automata can be obtained only by taking a broader view of these things and considering automata which can have outputs something like themselves. Now, one has to be careful what one means by this. There is no question of producing matter out of nothing. Rather, one imagines automata which can modify objects similar to themselves, or effect syntheses by picking up parts and putting them together, or take synthesized entities apart. In order to discuss these things, one has to imagine a formal set-up like this. Draw up a list of unambiguously defined elementary parts. Imagine that there is a practically unlimited supply of these parts floating around in a large container. One can then imagine an automaton functioning in the following manner: It also is floating around in this medium; its essential activity is to pick up parts and put them together, or, if aggregates of parts are found, to take them apart.

This is an axiomatically shortened and simplified description of

what an organism does. It's true that this view has certain limitations, but they are not fundamentally different from the inherent limitations of the axiomatic method. Any result one might reach in this manner will depend quite essentially on how one has chosen to define the elementary parts. It is a commonplace of all axiomatic methods that it is very difficult to give rigorous rules as to how one should choose the elementary parts, so that whether the choice of the elements was reasonable is a matter of common sense judgment. There is no rigorous description of what choice is reasonable and what choice is not.

First of all, one may define parts in such numbers, and each of them so large and involved, that one has defined the whole problem away. If you chose to define as elementary objects things which are analogous to whole living organisms, then you obviously have killed the problem, because you would have to attribute to these parts just those functions of the living organism which you would like to describe or to understand. So, by choosing the parts too large, by attributing too many and too complex functions to them, you lose the problem at the moment of defining it.

One also loses the problem by defining the parts too small, for instance, by insisting that nothing larger than a single molecule, single atom, or single elementary particle will rate as a part. In this case one would probably get completely bogged down in questions which, while very important and interesting, are entirely anterior to our problem. We are interested here in organizational questions about complicated organisms, and not in questions about the structure of matter or the quantum mechanical background of valency chemistry. So, it is clear that one has to use some common sense criteria about choosing the parts neither too large nor too small.

Even if one chooses the parts in the right order of magnitude, there are many ways of choosing them, none of which is intrinsically much better than any other. There is in formal logics a very similar difficulty, that the whole system requires an agreement on axioms, and that there are no rigorous rules on how axioms should be chosen, just the common sense rules that one would like to get the system one is interested in and would not like to state in his axioms either things which are really terminal theorems of his theory or things which belong to vastly anterior fields. For example, in axiomatizing geometry one should assume theorems from set theory, because one is not interested in how to get from sets to numbers, or from numbers to geometry. Again, one does not choose the more sophisticated theorems of analytic number theory as axioms of geometry, because one wants to cut in at an earlier point.

Even if the axioms are chosen within the common sense area, it is usually very difficult to achieve an agreement between two people who have done this independently. For instance, in the literature of formal logics there are about as many notations as there are authors, and anybody who has used a notation for a few weeks feels that it's more or less superior to any other. So, while the choice of notations, of the elements, is enormously important and absolutely basic for an application of the axiomatic method, this choice is neither rigorously justifiable nor humanly unambiguously justifiable. All one can do is to try to submit a system which will stand up under common sense criteria. I will give an indication of how one system can be constructed, but I want to emphasize very strongly how relatively I state this system.

I will introduce as elementary units neurons, a "muscle," entities which make and cut fixed contacts, and entities which supply energy, all defined with about that degree of superficiality with which the formal theory of McCulloch and Pitts describes an actual neuron. If you describe muscles, connective tissues, "disconnecting tissues," and means of providing metabolic energy, all with this degree of schematization, you wind up with a system of elements with which you can work in a reasonably uncomplicated manner. You probably wind up with something like 10 or 12 or 15 elementary parts.

By axiomatizing automata in this manner, one has thrown half of the problem out the window, and it may be the more important half. One has resigned oneself not to explain how these parts are made up of real things, specifically, how these parts are made up of actual elementary particles, or even of higher chemical molecules. One does not ask the most intriguing, exciting, and important question of why the molecules or aggregates which in nature really occur in these parts are the sort of things they are, why they are essentially very large molecules in some cases but large aggregations in other cases, why they always lie in a range beginning at a few microns and ending at a few decimeters. This is a very peculiar range for an elementary object, since it is, even on a linear scale, at least five powers of ten away from the sizes of really elementary entities.

These things will not be explained; we will simply assume that elementary parts with certain properties exist. The question that one can then hope to answer, or at least investigate, is: What principles are involved in organizing these elementary parts into functioning organisms, what are the traits of such organisms, and what are the essential quantitative characteristics of such organisms? I will discuss the matter entirely from this limited point of view.

[At this point von Neumann made the remarks on information, logic, thermodynamics, and balance which now appear at the end of the Third Lecture. They are placed there because that is where von Neumann's detailed outline located them. Those remarks are relevant to the present discussion because the concept of complication which von Neumann introduced next belongs to information theory.]

There is a concept which will be quite useful here, of which we have a certain intuitive idea, but which is vague, unscientific, and imperfect. This concept clearly belongs to the subject of information, and quasi-thermodynamical considerations are relevant to it. I know no adequate name for it, but it is best described by calling it "complication." It is effectivity in complication, or the potentiality to do things. I am not thinking about how involved the object is, but how involved its purposive operations are. In this sense, an object is of the highest degree of complexity if it can do very difficult and involved things.

I mention this because when you consider automata whose normal function is to synthesize other automata from elementary parts (living organisms and such familiar artificial automata as machine tools), you find the following remarkable thing. There are two states of mind, in each of which one can put himself in a minute, and in each of which we feel that a certain statement is obvious. But each of these two statements is the opposite or negation of the other!

Anybody who looks at living organisms knows perfectly well that they can produce other organisms like themselves. This is their normal function, they wouldn't exist if they didn't do this, and it's plausible that this is the reason why they abound in the world. In other words, living organisms are very complicated aggregations of elementary parts, and by any reasonable theory of probability or thermodynamics highly improbable. That they should occur in the world at all is a miracle of the first magnitude; the only thing which removes, or mitigates, this miracle is that they reproduce themselves. Therefore, if by any peculiar accident there should ever be one of them, from there on the rules of probability do not apply, and there will be many of them, at least if the milieu is reasonable. But a reasonable milieu is already a thermodynamically much less improbable thing. So, the operations of probability somehow leave a loophole at this point, and it is by the process of self-reproduction that they are pierced.

Furthermore, it's equally evident that what goes on is actually one degree better than self-reproduction, for organisms appear to have gotten more elaborate in the course of time. Today's organisms are phylogenetically descended from others which were vastly simpler

than they are, so much simpler, in fact, that it's inconceivable how any kind of description of the later, complex organism could have existed in the earlier one. It's not easy to imagine in what sense a gene, which is probably a low order affair, can contain a description of the human being which will come from it. But in this case you can say that since the gene has its effect only within another human organism, it probably need not contain a complete description of what is to happen, but only a few cues for a few alternatives. However, this is not so in phylogenetic evolution. That starts from simple entities, surrounded by an unliving amorphous milieu, and produces something more complicated. Evidently, these organisms have the ability to produce something more complicated than themselves.

The other line of argument, which leads to the opposite conclusion, arises from looking at artificial automata. Everyone knows that a machine tool is more complicated than the elements which can be made with it, and that, generally speaking, an automaton A, which can make an automaton B, must contain a complete description of B and also rules on how to behave while effecting the synthesis. So, one gets a very strong impression that complication, or productive potentiality in an organization, is degenerative, that an organization which synthesizes something is necessarily more complicated, of a higher order, than the organization it synthesizes. This conclusion, arrived at by considering artificial automata, is clearly opposite to our early conclusion, arrived at by considering living organisms.

I think that some relatively simple combinatorial discussions of artificial automata can contribute to mitigating this dilemma. Appealing to the organic, living world does not help us greatly, because we do not understand enough about how natural organisms function. We will stick to automata which we know completely because we made them, either actual artificial automata or paper automata described completely by some finite set of logical axioms. It is possible in this domain to describe automata which can reproduce themselves. So at least one can show that on the site where one would expect complication to be degenerative it is not necessarily degenerative at all, and, in fact, the production of a more complicated object from a less complicated object is possible.

The conclusion one should draw from this is that complication is degenerative below a certain minimum level. This conclusion is quite in harmony with other results in formal logics, to which I have referred a few times earlier during these lectures.[1] We do not now know

[1] [See the end of the Second Lecture.]

what complication is, or how to measure it, but I think that something like this conclusion is true even if one measures complication by the crudest possible standard, the number of elementary parts. There is a minimum number of parts below which complication is degenerative, in the sense that if one automaton makes another the second is less complex than the first, but above which it is possible for an automaton to construct other automata of equal or higher complexity. Where this number lies depends upon how you define the parts. I think that with reasonable definitions of parts, like those I will partially indicate later, which give one or two dozen parts with simple properties, this minimum number is large, in the millions. I don't have a good estimate of it, although I think that one will be produced before terribly long, but to do so will be laborious.

There is thus this completely decisive property of complexity, that there exists a critical size below which the process of synthesis is degenerative, but above which the phenomenon of synthesis, if properly arranged, can become explosive, in other words, where syntheses of automata can proceed in such a manner that each automaton will produce other automata which are more complex and of higher potentialities than itself.

Now, none of this can get out of the realm of vague statement until one has defined the concept of complication correctly. And one cannot define the concept of complication correctly until one has seen in greater detail some critical examples, that is, some of the constructs which exhibit the critical and paradoxical properties of complication. There is nothing new about this. It was exactly the same with conservation and non-conservation properties in physics, with the concepts of energy and entropy, and with other critical concepts. The simplest mechanical and thermodynamic systems had to be discussed for a long time before the correct concepts of energy and entropy could be abstracted from them.

[Von Neumann only briefly described the kinds of elements or parts he planned to use. There are neurons like those of McCulloch and Pitts. There are elements "that have absolutely no function except that they are rigid and produce a geometrical tie between their ends." Another kind of element is called a "motor organ" and a "muscle-like affair"; it contracts to zero length when stimulated. There is an organ which, when pulsed, "can either make or break a connection." He said that less than a dozen kinds of elements are needed. An automaton composed of these parts can catch other parts which accidentally come in contact with it; "it is possible to invent a system by which it can sense" what part it has caught.

In June of 1948 von Neumann gave three lectures on automata at the Institute for Advanced Study to a small group of friends. He probably did this in preparation for the Hixon Symposium which took place in September of that year.[2] These lectures contained the most detailed description of the parts of his self-reproducing automaton that I know of. For this reason, I have attempted to reconstruct, from the notes and memories of the audience, what he said about these parts and how they would function.

Von Neumann described eight kinds of parts. All seem to have been symbolized with straight lines; inputs and outputs were indicated at the ends and/or the middle. The temporal reference frame was discrete, each element taking a unit of time to respond. It is not clear whether he intended this list to be complete; I suspect that he had not yet made up his mind on this point.

Four of the parts perform logical and information processing operations. A *stimulus organ* receives and transmits stimuli; it receives them disjunctively, that is, it realizes the truth-function "p or q." A *coincidence organ* realizes the truth-function "p and q." An *inhibitory organ* realizes the truth-function "p and not-q." A *stimuli producer* serves as a source of stimuli.

The fifth part is a *rigid member*, from which a rigid frame for an automaton can be constructed. A rigid member does not carry any stimuli; that is, it is an insulated girder. A rigid member may be connected to other rigid members as well as to parts which are not rigid members. These connections are made by a *fusing organ* which, when stimulated, welds or solders two parts together. Presumably the fusing organ is used in the following way. Suppose point a of one girder is to be joined to point b of another girder. The active or output end of the fusing organ is placed in contact with points a and b. A stimulus into the input end of the fusing organ at time t causes points a and b to be welded together at time $t + 1$. The fusing organ can be withdrawn later. Connections may be broken by a *cutting organ* which, when stimulated, unsolders a connection.

The eighth part is a *muscle*, used to produce motion. A muscle is normally rigid. It may be connected to other parts. If stimulated at time t it will contract to length zero by time $t + 1$, keeping all its connections. It will remain contracted as long as it is stimulated. Presumably muscles can be used to move parts and make connections in the following way. Suppose that muscle 1 lies between point a of

[2] ["The General and Logical Theory of Automata." *Collected Works* 5.288–328. It will be recalled that the Illinois lectures were delivered in December of 1949.]

one girder and point b of another girder, and muscle 2 lies between point a and the active end c of a fusing organ. When both muscles are stimulated, they will contract, thereby bringing points a, b, and c together. When the fusing organ is stimulated, it will weld points a and b together. Finally, when the stimuli to the muscles are stopped, the muscles will return to their original length, at least one end of muscle 1 separating from the point ab. Von Neumann does not seem to have discussed the question of how the connections between muscles and other parts are made and broken.

Von Neumann conceived of an automaton constructing other automata in the following manner. The constructing automaton floats on a surface, surrounded by an unlimited supply of parts. The constructing automaton contains in its memory a description of the automaton to be constructed. Operating under the direction of this description, it picks up the parts it needs and assembles them into the desired automaton. To do this, it must contain a device which catches and identifies the parts that come in contact with it. The June, 1948 lectures contain only a few remarks on how this device might operate. Two stimulus units protrude from the constructing automaton. When a part touches them tests can be made to see what kind of part it is. For example, a stimulus organ will transmit a signal; a girder will not. A muscle might be identified by determining that it contracts when stimulated.

Von Neumann intended to disregard the fuel and energy problem in his first design attempt. He planned to consider it later, perhaps by introducing a battery as an additional elementary part. Except for this addition, von Neumann's early model of self-reproduction deals with the geometrical-kinematic problems of movement, contact, positioning, fusing, and cutting, and ignores the truly mechanical and chemical questions of force and energy. Hence I call it his *kinematic model* of self-reproduction. This early model is to be contrasted with his later *cellular model* of self-reproduction, which is presented in Part II of the present work.

In his June, 1948 lectures von Neumann raised the question of whether kinematic self-reproduction requires three dimensions. He suspected that either three dimensions or a Riemann surface (multiply-connected plane) would be needed. We will see in Part II that only two dimensions are required for self-reproduction in von Neumann's cellular model. This is a strong indication that two dimensions are sufficient for kinematic self-reproduction.

We return now to the Illinois lectures. Von Neumann discussed the general design of a self-reproducing automaton. He said that it

is in principle possible to set up a machine shop which can make a copy of any machine, given enough time and raw materials. This shop would contain a machine tool B with the following powers. Given a pattern or object X, it would search over X and list its parts and their connections, thereby obtaining a description of X. Using this description, the tool B would then make a copy of X. "This is quite close to self-reproduction, because you can furnish B with itself."]

But it is easier, and for the ultimate purpose just as effective, not to construct an automaton which can copy any pattern or specimen given to it, but to construct an automaton which can produce an object starting from a logical description. In any conceivable method ever invented by man, an automaton which produces an object by copying a pattern will go first from the pattern to a description and then from the description to the object. It first abstracts what the thing is like, and then carries it out. It's therefore simpler not to extract from a real object its definition, but to start from the definition.

To proceed in this manner one must have axiomatic descriptions of automata. You see, I'm coming quite close to Turing's trick with universal automata, which also started with a general formal description of automata. If you take those dozen elements I referred to in a rather vague and general way and give exact descriptions of them (which could be done on two printed pages or less), you will have a formal language for describing automata unambiguously. Now any notation can be expressed as a binary notation, which can be recorded on a punched tape with a single channel. Hence any automaton description could be punched on a piece of tape. At first, it is better not to use a description of the pieces and how they fit together, but rather a description of the consecutive steps to be used in building the automaton.

[Von Neumann then showed how to construct a binary tape out of rigid elements. See Figure 2. A binary character is represented at each intersection of the basic chain; "one" is represented by an attached rigid element, "zero" by the absence of a side element. Writing and erasing are accomplished by adding and removing side elements.]

I have simplified unnecessarily, just because of a purely mathematical habit of trying to do things with a minimum of notation. Since I'm using a binary notation, all I'm attaching here is no side chain, or a one-step side chain. Existing languages and practical notations use more symbols than the binary system. There is no difficulty in using more symbols here; you simply attach more complex side chains. In fact, the very linearity of our logical notation is

completely unnecessary here. You could use more complicated looped chains, which would be perfectly good carriers for a code, but it would not be a linear code. There is reason to suspect that our predilection for linear codes, which have a simple, almost temporal sequence, is chiefly a literary habit, corresponding to our not particularly high level of combinatorial cleverness, and that a very efficient language would probably depart from linearity.[3]

There is no great difficulty in giving a complete axiomatic account of how to describe any conceivable automaton in a binary code. Any such description can then be represented by a chain of rigid elements like that of Figure 2. Given any automaton X, let $\phi(X)$ designate the chain which represents X. Once you have done this, you can design a universal machine tool A which, when furnished with such a chain $\phi(X)$, will take it and gradually consume it, at the same time building up the automaton X from the parts floating around freely in the surrounding milieu. All this design is laborious, but it is not difficult in principle, for it's a succession of steps in formal logics. It is not qualitatively different from the type of argumentation with which Turing constructed his universal automaton.

Another thing which one needs is this. I stated earlier that it might be quite complicated to construct a machine which will copy an automaton that is given it, and that it is preferable to proceed, not from original to copy, but from verbal description to copy. I would like to make one exception; I would like to be able to copy linear chains of rigid elements. Now this is very easy. For the real reason it is harder to copy an existing automaton than its description is that the existing automaton does not conform with our habit of linearity, its parts being connected with each other in all possible directions, and it's quite difficult just to check off the pieces that have already been described.[4] But it's not difficult to copy a linear chain of rigid elements. So I will assume that there exists an automaton B which has this property: If you provide B with a description of anything, it consumes it and produces two copies of this description.

Please consider that after I have described these two elementary steps, one may still hold the illusion that I have not broken the principle of the degeneracy of complication. It is still not true that, starting from something, I have made something more subtle and more

[3] [The programming language of flow diagrams, invented by von Neumann, is a possible example. See p. 13 of the Introduction to the present volume.]

[4] [Compare Sec. 1.6.3 of Part II, written about 3 years later. Here von Neumann gives a more fundamental reason for having the constructing automaton work from a description of an automaton rather than from the automaton itself.]

involved. The general constructive automaton A produces only X when a complete description of X is furnished it, and on any reasonable view of what constitutes complexity, this description of X is as complex as X itself. The general copying automaton B produces two copies of $\phi(X)$, but the juxtaposition of two copies of the same thing is in no sense of higher order than the thing itself. Furthermore, the extra unit B is required for this copying.

Now we can do the following thing. We can add a certain amount of control equipment C to the automaton $A + B$. The automaton C dominates both A and B, actuating them alternately according to the following pattern. The control C will first cause B to make two copies of $\phi(X)$. The control C will next cause A to construct X at the price of destroying one copy of $\phi(X)$. Finally, the control C will tie X and the remaining copy of $\phi(X)$ together and cut them loose from the complex $(A + B + C)$. At the end the entity $X + \phi(X)$ has been produced.

Now choose the aggregate $(A + B + C)$ for X. The automaton $(A + B + C) + \phi(A + B + C)$ will produce $(A + B + C) + \phi(A + B + C)$. Hence auto-reproduction has taken place.

[The details are as follows. We are given the universal constructor $(A + B + C)$, to which is attached a description of itself, $\phi(A + B + C)$. Thus the process of self-reproduction starts with $(A + B + C) + \phi(A + B + C)$. Control C directs B to copy the description twice; the result is $(A + B + C) + \phi(A + B + C) + \phi(A + B + C)$. Then C directs A to produce the automaton $A + B + C$ from one copy of the description; the result is $(A + B + C) + (A + B + C) + \phi(A + B + C)$. Finally, C ties the new automaton and its description together and cuts them loose. The final result consists of the two automata $(A + B + C)$ and $(A + B + C) + \phi(A + B + C)$. If B were to copy the description thrice, the process would start with one copy of $(A + B + C) + \phi(A + B + C)$ and terminate with two copies of this automaton. In this way, the universal constructor reproduces itself.]

This is not a vicious circle. It is quite true that I argued with a variable X first, describing what C is supposed to do, and then put something which involved C for X. But I defined A and B exactly, before I ever mentioned this particular X, and I defined C in terms which apply to any X. Therefore, in defining A, B, and C, I did not make use of what X is to be, and I am entitled later on to use an X which refers explicitly to A, B, and C. The process is not circular.

The general constructive automaton A has a certain creative ability, the ability to go from a description of an object to the object. Like-

wise, the general copying automaton B has the creative ability to go from an object to two copies of it. Neither of these automata, however, is self-reproductive. Moreover, the control automaton C is far from having any kind of creative or reproductive ability. All it can do is to stimulate two other organs so that they act in certain ways, tie certain things together, and cut these things loose from the original system. Yet the combination of the three automata A, B, and C is auto-reproductive. Thus you may break a self-reproductive system into parts whose functioning is necessary for the whole system to be self-reproductive, but which are not themselves self-reproductive.

You can do one more thing. Let X be $A + B + C + D$, where D is any automaton. Then $(A + B + C) + \phi(A + B + C + D)$ produces $(A + B + C + D) + \phi(A + B + C + D)$. In other words, our constructing automaton is now of such a nature that in its normal operation it produces another object D as well as making a copy of itself. This is the normal function of an auto-reproductive organism: it creates byproducts in addition to reproducing itself.

The system $(A + B + C + D)$ can undergo processes similar to the process of mutation. One of the difficulties in defining what one means by self-reproduction is that certain organizations, such as growing crystals, are self-reproductive by any naive definition of self-reproduction, yet nobody is willing to award them the distinction of being self-reproductive. A way around this difficulty is to say that self-reproduction includes the ability to undergo inheritable mutations as well as the ability to make another organism like the original.

Consider the situation with respect to the automaton $(A + B + C + D) + \phi(A + B + C + D)$. By a mutation I will simply mean a random change of one element anywhere. If an element is changed at random in one of the automata A, B, or C, the system will usually not completely reproduce itself. For example, if an element is changed in C, C may fail to stimulate A and B at the proper time, or it may fail to make the connections and disconnections which are required. Such a mutation is lethal.

If there is a change in the description $\phi(A + B + C + D)$, the system will produce, not itself, but a modification of itself. Whether the next generation can produce anything or not depends on where the change is. If the change is in A, B, or C, the next generation will be sterile. If the change occurs in D, the system with the mutation is exactly like the original system, except that D has been replaced by D'. This system can reproduce itself, but its by-product will be

D' rather than D. This is the normal pattern of an inheritable mutation.

So, while this system is exceedingly primitive, it has the trait of an inheritable mutation, even to the point that a mutation made at random is most probably lethal, but may be non-lethal and inheritable.