Some Moral and Technical Consequences of Automation

As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

Norbert Wiener

Some 13 years ago, a book of mine was published by the name of *Cybernetics.* In it I discussed the problems of control and communication in the living organism and the machine. I made a considerable number of predictions about the development of controlled machines and about the corresponding techniques of automatization, which I foresaw as having important consequences affecting the society of the future. Now, 13 years later, it seems appropriate to take stock of the present position with respect to both cybernetic technique and the social consequences of this technique.

Before commencing on the detail of these matters, I should like to mention a certain attitude of the man in the street toward cybernetics and automatization. This attitude needs a critical discussion, and in my opinion it should be rejected in its entirety. This is the assumption that machines cannot possess any degree of originality. This frequently takes the form of a statement that nothing can come out of the machine which has not been put into it. This is often interpreted as asserting that a machine which man has made must remain continually subject to man, so that its operation is at any time open to human interference and to a change in policy. On the basis of such an attitude, many people have pooh-poohed the dangers of machine techniques, and they have flatly contradicted the early predictions of Samuel Butler that the machine might take over the control of mankind.

It is true that in the time of Samuel Butler the available machines were far less hazardous than machines are today, for they involved only power, not a certain degree of thinking and communication. However, the machine techniques of the present day have invaded the latter fields as well, so that the actual machine of today is very different from the image that Butler held, and we cannot transfer to these new devices the assumptions which seemed axiomatic a generation ago. I find myself facing a public which has formed its attitude toward the machine on the basis of an imperfect understanding of the structure and mode of operation of modern machines.

It is my thesis that machines can and do transcend some of the limitations of their designers, and that in doing so they may be both effective and dangerous. It may well be that in principle we cannot make any machine the elements of whose behavior we cannot comprehend sooner or later. This does not mean in any way that we shall be able to comprehend these elements in substantially less time than the time required for operation of the machine, or even within any given number of years or generations.

As is now generally admitted, over a limited range of operation, machines act far more rapidly than human beings and are far more precise in performing the details of their operations. This being the case, even when machines do not in any way transcend man's intelligence, they very well may, and often do, transcend man in the performance of tasks. An intelligent understanding of their mode of performance may be delayed until long after the task which they have been set has been completed.

This means that though machines are theoretically subject to human criticism, such criticism may be ineffective until long after it is relevant. To be effective in warding off disastrous consequences, our understanding of our man-made machines should in general develop *pari passu* with the performance of the machine. By the very slowness of our human actions, our effective control of our machines may be nullified. By the time we are able to react to information conveyed by our senses and stop the car we are driving, it may already have run head on into a wall.

Game-Playing

I shall come back to this point later in this article. For the present, let me discuss the technique of machines for a very specific purpose: that of playing games. In this matter I shall deal more particularly with the game of checkers, for which the International Business Machines Corporation has developed very effective game-playing machines.

Let me say once for all that we are not concerned here with the machines which operate on a perfect closed theory of the game they play. The game theory of von Neumann and Morgenstern may be suggestive as to the operation of actual game-playing machines, but it does not actually describe them.

In a game as complicated as checkers, if each player tries to choose his play in view of the best move his opponent can make, against the best response he can give, against the best response his opponent can give, and so on, he will have taken upon himself an impossible task. Not only is this humanly impossible but there is actually no reason to suppose that it is the best policy against the opponent by whom he is faced, whose limitations are equal to his own.

The von Neumann theory of games bears no very close relation to the theory by which game-playing machines operate. The latter corresponds much more closely to the methods of play used by expert but limited human chess players against other chess players. Such players depend on certain strategic evaluations, which are in essence not complete. While the von Neumann type of play is valid for games like ticktacktoe, with a complete theory, the very interest of chess and checkers lies in the fact that they

The author is professor of mathematics at Massachusetts Institute of Technology, Cambridge. This article is adapted from a lecture he delivered 27 December 1959 before the Committee on Science in the Promotion of Human Welfare, at the Chicago meeting of the AAAS.

do not possess a complete theory. Neither do war, nor business competition, nor any of the other forms of competitive activity in which we are really interested.

In a game like ticktacktoe, with a small number of moves, where each player is in a position to contemplate all possibilities and to establish a defense against the best possible moves of the other player, a complete theory of the von Neumann type is valid. In such a case, the game must inevitably end in a win for the first player, a win for the second player, or a draw.

I question strongly whether this concept of the perfect game is a completely realistic one in the cases of actual, nontrivial games. Great generals like Napoleon and great admirals like Nelson have proceeded in a different manner. They have been aware not only of the limitations of their opponents in such matters as materiel and personnel but equally of their limitations in experience and in military know-how. It was by a realistic appraisal of the relative inexperience in naval operations of the continental powers as compared with the highly developed tactical and strategic competence of the British fleet that Nelson was able to display the boldness which pushed the continental forces off the seas. This he could not have done had he engaged in the long, relatively indecisive, and possibly losing conflict to which his assumption of the best possible strategy on the part of his enemy would have doomed him.

In assessing not merely the materiel and personnel of his enemies but also the degree of judgment and the amount of skill in tactics and strategy to be expected of them, Nelson acted on the basis of their record in previous combats. Similarly, an important factor in Napoleon's conduct of his combat with the Austrians in Italy was his knowledge of the rigidity and mental limitations of Würmser.

This element of experience should receive adequate recognition in any realistic theory of games. It is quite legitimate for a chess player to play, not against an ideal, nonexisting, perfect antagonist, but rather against one whose habits he has been able to determine from the record. Thus, in the theory of games, at least two different intellectual efforts must be made. One is the short-term effort of playing with a determined policy for the individual game. The other is the examination of a record of many games. This record has been set by the player himself, by his opponent, or even by players with whom he has not personally played. In terms of this record, he determines the relative advantages of different policies as proved over the past.

There is even a third stage of judgment required in a chess game. This is expressed at least in part by the length of the significant past. The development of theory in chess decreases the importance of games played at a different stage of the art. On the other hand, an astute chess theoretician may estimate in advance that a certain policy currently in fashion has become of little value, and that it may be best to return to earlier modes of play to anticipate the change in policy of the people whom he is likely to find as his opponents.

Thus, in determining policy in chess there are several different levels of consideration which correspond in a certain way to the different logical types of Bertrand Russell. There is the level of tactics, the level of strategy, the level of the general considerations which should have been weighed in determining this strategy, the level in which the length of the relevant past—the past within which these considerations may be valid—is taken into account, and so on. Each new level demands a study of a much larger past than the previous one.

I have compared these levels with the logical types of Russell concerning classes, classes of classes, classes of classes of classes, and so on. It may be noted that Russell does not consider statements involving all types as significant. He brings out the futility of such questions as that concerning the barber who shaves all persons, and only those persons, who do not shave themselves. Does he shave himself? On one type he does, on the next type he does not, and so on, indefinitely. All such questions involving an infinity of types may lead to unsolvable paradoxes. Similarly, the search for the best policy under all levels of sophistication is a futile one and must lead to nothing but confusion.

These considerations arise in the determination of policy by machines as well as in the determination of policy by persons. These are the questions which arise in the programming of programming. The lowest type of game-playing machine plays in terms of a certain rigid evaluation of plays. Quantities such as the value of pieces gained or lost, the command of the pieces, their mobility, and so on, can be given numerical weights on a certain empirical basis, and a weighting may be given on this basis to each next play conforming to the rules of the game. The play with the greatest weight may be chosen. Under these circumstances, the play of the machine will seem to its antagonist—who cannot help but evaluate the chess personality of the machine—a rigid one.

Learning Machines

The next step is for the machine to take into consideration not merely the moves as they occurred in the individual game but the record of games previously played. On this basis, the machine may stop from time to time, not to play but to consider what (linear or nonlinear) weighting of the factors which it has been given to consider would correspond best to won games as opposed to lost (or drawn) games. On this basis, it continues to play with a new weighting. Such a machine would seem to its human opponent to have a far less rigid game personality, and tricks which would defeat it at an earlier stage may now fail to deceive it.

The present level of these learning machines is that they play a fair amateur game at chess but that in checkers they can show a marked superiority to the player who has programmed them after from 10 to 20 playing hours of working and indoctrination. They thus most definitely escape from the completely effective control of the man who has made them. Rigid as the repertory of factors may be which they are in a position to take into consideration, they do unquestionably—and so say those who have played with them—show originality, not merely in their tactics, which may be quite unforeseen, but even in the detailed weighting of their strategy.

As I have said, checker-playing machines which learn have developed to the point at which they can defeat the programmer. However, they appear still to have one weakness. This lies in the end game. Here the machines are somewhat clumsy in determining the best way to give the *coup de grâce*. This is due to the fact that the existing machines have for the most part adopted a program in which the identical strategy is carried out at each stage of the game. In view of the similarity of values of pieces in checkers, this is quite natural for a large part of the play but ceases to be perfectly relevant when the board is relatively empty and the main problem is that of moving into position rather than that of direct attack. Within the frame of the methods I have described it is quite possible to have a second exploration to determine what the policy should be after the number of pieces of the opponent is so reduced that these new considerations become paramount.

Chess-playing machines have not, so far, been brought to the degree of perfection of checker-playing machines, although, as I have said, they can most certainly play a respectable amateur game. Probably the reason for this is similar to the reason for their relative efficiency in the end game of checkers. In chess, not only is the end game quite different in its proper strategy from the mid-game but the opening game is also. The difference between checkers and chess in this respect is that the initial play of the pieces in checkers is not very different in character from the play which arises in the mid-game, while in chess, pieces at the beginning have an arrangement of exceptionally low mobility, so that the problem of deploying them from this position is particularly difficult. This is the reason why opening play and development form a special branch of chess theory.

There are various ways in which the machine can take cognizance of these well-known facts and explore a separate waiting strategy for the opening. This does not mean that the type of game theory which I have here discussed is not applicable to chess but merely that it requires much more consideration before we can make a machine that can play master chess. Some of my friends who are engaged in these problems believe that this goal will be achieved in from 10 to 25 years. Not being a chess expert, I do not venture to make any such predictions on my own initiative.

It is quite in the cards that learning machines will be used to program the pushing of the button in a new push-button war. Here we are considering a field in which automata of a non-learning character are probably already in use. It is quite out of the question to program these machines on the basis of an actual experience in real war. For one thing, a sufficient experience to give an adequate programming would probably see humanity already wiped out.

Moreover, the techniques of push-button war are bound to change so much that by the time an adequate experience could have been accumulated, the basis of the beginning would have radically changed. Therefore, the programming of such a learning machine would have to be based on some sort of war game, just as commanders and staff officials now learn an important part of the art of strategy in a similar manner. Here, however, if the rules for victory in a war game do not correspond to what we actually wish for our country, it is more than likely that such a machine may produce a policy which would win a nominal victory on points at the cost of every interest we have at heart, even that of national survival.

Man and Slave

The problem, and it is a moral problem, with which we are here faced is very close to one of the great problems of slavery. Let us grant that slavery is bad because it is cruel. It is, however, self-contradictory, and for a reason which is quite different. We wish a slave to be intelligent, to be able to assist us in the carrying out of our tasks. However, we also wish him to be subservient. Complete subservience and complete intelligence do not go together. How often in ancient times the clever Greek philosopher slave of a less intelligent Roman slaveholder must have dominated the actions of his master rather than obeyed his wishes! Similarly, if the machines become more and more efficient and operate at a higher and higher psychological level, the catastrophe foreseen by Butler of the dominance of the machine comes nearer and nearer.

The human brain is a far more efficient control apparatus than is the intelligent machine when we come to the higher areas of logic. It is a self-organizing system which depends on its capacity to modify itself into a new machine rather than on ironclad accuracy and speed in problem-solving. We have already made very successful machines of the lowest logical type, with a rigid policy. We are beginning to make machines of the second logical type, where the policy itself improves with learning. In the construction of operative machines, there is no specific foreseeable limit with respect to logical type, nor is it safe to make a pronouncement about the exact level at which the brain is superior to the machine. Yet for a long time at least there will always be some level at which the brain is better than the constructed machine, even though this level may shift upwards and upwards.

It may be seen that the result of a programming technique of automatization is to remove from the mind of the designer and operator an effective understanding of many of the stages by which the machine comes to its conclusions and of what the real tactical intentions of many of its operations may be. This is highly relevant to the problem of our being able to foresee undesired consequences outside the frame of the strategy of the game while the machine is still in action and while intervention on our part may prevent the occurrence of these consequences.

Here it is necessary to realize that human action is a feedback action. To avoid a disastrous consequence, it is not enough that some action on our part should be sufficient to change the course of the machine, because it is quite possible that we lack information on which to base consideration of such an action.

In neurophysiological language, ataxia can be quite as much of a deprivation as paralysis. A patient with locomotor ataxia may not suffer from any defect of his muscles or motor nerves, but if his muscles and tendons and organs do not tell him exactly what position he is in, and whether the tensions to which his organs are subjected will or will not lead to his falling, he will be unable to stand up. Similarly, when a machine constructed by us is capable of operating on its incoming data at a pace which we cannot keep, we may not know, until too late, when to turn it off. We all know the fable of the sorcerer's apprentice, in which the boy makes the broom carry water in his master's absence, so that it is on the point of drowning him when his master reappears. If the boy had had to seek a charm to stop the mischief in the *grimoires* of his master's library, he might have been drowned before he had discovered the relevant incantation. Similarly, if a bottle factory is programmed on the basis of maximum productivity, the owner may be made bankrupt by the enormous inventory of unsalable bottles manufactured before he learns he should have stopped production six months earlier.

The "Sorcerer's Apprentice" is only one of many tales based on the assumption that the agencies of magic are literal-minded. There is the story of the genie and the fisherman in the *Arabian Nights,* in which the fisherman breaks the seal of Solomon which has imprisoned the genie and finds the genie vowed to his own destruction; there is the tale of the "Monkey's Paw," by W. W. Jacobs, in which the sergeant major brings back from India a talisman which has the power to grant each of three people three wishes. Of the first recipient of this talisman we are told only that his third wish is for death. The sergeant major, the second person whose wishes are granted, finds his experiences too terrible to relate. His friend, who receives the talisman, wishes first for £200. Shortly thereafter, an official of the factory in which his son works comes to tell him that his son has been killed in the machinery and that, without any admission of responsibility, the company is sending him as consolation the sum of £200. His next wish is that his son should come back, and the ghost knocks at the door. His third wish is that the ghost should go away.

Disastrous results are to be expected not merely in the world of fairy tales but in the real world wherever two agencies essentially foreign to each other are coupled in the attempt to achieve a common purpose. If the communication between these two agencies as to the nature of this purpose is incomplete, it must only be expected that the results of this cooperation will be unsatisfactory. If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it, because the action is so fast and irrevocable that we have not the data to intervene before the action is complete, then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.

Time Scales

Up to this point I have been considering the quasi-moral problems caused by the simultaneous action of the machine and the human being in a joint enterprise. We have seen that one of the chief causes of the danger of disastrous consequences in the use of the learning machine is that man and machine operate on two distinct time scales, so that the machine is much faster than man and the two do not gear together without serious difficulties. Problems of the same sort arise whenever two control operators on very different time scales act together, irrespective of which system is the faster and which system is the slower. This leaves us the much more directly moral question: What are the moral problems when man as an individual operates in connection with the controlled process of a much slower time scale, such as a portion of political history or—our main subject of inquiry—the development of science?

Let it be noted that the development of science is a control and communication process for the long-term understanding and control of matter. In this process 50 years are as a day in the life of the individual. For this reason, the individual scientist must work as a part of a process whose time scale is so long that he himself can only contemplate a very limited sector of it. Here, too, communication between the two parts of a double machine is difficult and limited. Even when the individual believes that science contributes to the human ends which he has at heart, his belief needs a continual scanning and re-evaluation which is only partly possible. For the individual scientist, even the partial appraisal of this liaison between the man and the process requires an imaginative forward glance at history which is difficult, exacting, and only limitedly achievable. And if we adhere simply to the creed of the scientist, that an incomplete knowledge of the world and of ourselves is better than no knowledge, we can still by no means always justify the naive assumption that the faster we rush ahead to employ the new powers for action which are opened up to us, the better it will be. We must always exert the full strength of our imagination to examine where the full use of our new modalities may lead us.

Science in the News

The Jackson Committee: Educating the Next President and the Next Congress

The most civilized, and perhaps the most important, current congressional investigation is that being conducted by Sen. Henry Jackson (D-Wash.) and his Subcommittee on National Policy Machinery. Its purpose, in part, is the unusual one of educating the next president to the pitfalls involved in organizing his bewilderingly complex job.

The committee also hopes to develop legislation, where legislation might be helpful, to smooth the president's problem. Perhaps more important, the committee hopes to build a case for reorganizing certain procedures, particularly in the area of the budget, which clearly need alteration, but which are likely to remain unchanged until basic attitudes in Congress are gradually changed.

James Reston, of the New York *Times*, has described the committee's efforts as "legislative investigation at its very best . . . scholarly, objective and nonpartisan." A measure of Jackson's success in meeting these refreshing standards is that the minority counsel, present to see that the witnesses put on record their estimates of the strong as well as weak points of the administration, has very little to do. This has not been because the committee has failed so far to uncover any areas of weakness, but because the committee has so