vordenker
Winter-Edition 2001

TOWARD SOME CIRCUITRY OF ETHICAL ROBOTS

or

An Observational Science of the Genesis of Social Evaluation in the Mind-Like Behavior of Artifacts

by W. S. McCulloch

Read to the 13th Conference on Science, Philosophy, and Religion, New York, September 1952, and to the Meeting under the Auspices of the Department of Experimental Psychiatry, University of Birmingham, England, 1953.

Argument. No empiricist expects to find in men or machines exceptions to natural law, but physical sciences are not constructed to state or solve those problems of biology, psychology or sociology that involve adaptive, perceptive, thoughtful or communicative behavior. Recently, two new sciences have arisen in which these problems may be stated. Information theory, concerned with the amount of information carried by signals in the presence of noise, distinguishes between signals, which are true or else false, and noise, which is neither – a distinction alien to physics but crucial in the design of communication systems and applicable to all transducers of information and all computers, be they men or machines. Cybernetics stems from WIENER's insight into governors and servo systems: In the negative feedback by which their output decreases their input what has to return is not a physical affair, such as energy, but simply information on the outcome of a previous act. This new science has yielded a theory of all homeostatic and purposive behavior, regardless of the physical nature of the components subserving the circuit-action and, hence, is as applicable to men as to machines. Hundreds of articles attest its use in biology (STUMPERS). This is a modest attempt to show that these two new sciences have a bearing on social problems.

In conformity with English usage the word "ethics" in this article, denotes the character or mode of behavior that develops in social intercourse and serves the ends created by that association. My model of society is the smallest that can share in such an end; namely, two associates. Three, as VON NEUMANN's theory of games clearly indicates, permits any two to combine against the third, and larger numbers yield greater complexities. With three in a democracy, each preferring himself to the neighbor on his right, and him to the one on his left, which will be elected when only two are nominated clearly depends in a circular fashion upon which two are nominated. While we already know that in controlled situations the dynamics of three and larger groups depends upon what channels of communication are allowed, we are as yet ignorant of the nature of that dependence and unable to predict the outcome. Hence, what I have said here should not be mistaken for an analysis of a social situation involving more than two that share.

Also, I have used "moral" in a familiar English sense, for those modes of behavior that conform to instruction in, or to revelation of, the laws of right conduct. Thus, having heard a fable, we ask "What is the moral of that story?" For I desire to distinguish both the genesis of values from particular experiences and the acceptance of one set of values rather than another from merely instinctive or inherited sets of values, such as insects are thought to enjoy. If for these distinctions the reader finds more convenient terms, I shall be happy to employ them.

But in my simple model I have supposed too much. Clearly, given a desire to play and an opponent who quits when the beginner makes too many losing moves, the novice will learn what it is to win and will play to that end.

Lest you be misled, kindly remember that on questions of good and evil, science has nothing to say. But whether or not man can conceive a tautological theory of the good, like mathematics and logic, I mean a normative science of values, he can construct an observational science of evaluation. He must watch the choices of the organisms or machines to discover the causes of such conduct. But to be ethical, these must include other organisms and machines which must share effort and reward or no social questions of good and evil will arise. I shall investigate what Machines, by cooperation and competition, can constitute a society where their conduct becomes self-disciplined in a way that serves the ends created by their association. Two developments in the theory of machines and of information will serve us to begin.

The first of these, though old in the art of engineering, is new in the form that WIENER, ROSENBLUETH, and BIGELOW suggested in their paper on teleological mechanisms. There are many mechanical and electrical devices operating according to the principles laid down by them. In all of them, a change in the input produces an output which acts upon the input so as to diminish that change. All closed paths of this kind establish a certain state of the system as the end of its operation. They cause the device to return to that state whenever the world jogs him away from it, and are hence said to be "error operated". Governors and regulators are usually of this type. Those in the brain include some like the automatic volume control of a standard radio receiver. Certain reflexes, involving both the brain and the body, require circuits like the ones in the "power pack", which takes the variable alternating current from the line and supplies direct current at the constant voltage required by the tubes. But the circuits that are of principal importance here traverse the organism or the animal and its environment. They are appetitive, being "error operated" from some target, or goal, in the environment. Without them neither a man nor a machine would have purposes beyond his internal rearrangement. In terms of such circuits, when they traverse other men or other appetitive machines, we can conceive the purposes that engender ethics.

The mere fact that his fellows are appetitive, requires the machine to treat them as appetitive, even if he only wants to use them for or his own ends. This falls short of the Categorical Imperative; but may yet prove sufficient basis for an ethic of enlightened selfishness. We shall return presently to the requisite enlightenment and its mechanistic foundation, but for the moment we pause to consider what is sometimes called the "value anomaly"' (MCCULLOCH, 1945). By this we mean that an animal or machine, successively offered his choice between each two of three incompatible ends, A, B, C, sometimes chooses A rather than B, B rather than C, C rather than A, and does so consistently. I have myself encountered this in experimental esthetics, when examining by paired comparison three rectangles divided into 2, 3, and 5 equal rectangles. Animal psychologists have discovered that, say, a hundred male rats all deprived of food and sex for a specified period will all prefer food to sex, sex to, avoidance of shock, and avoidance of shock to food. That this happens is of theoretical importance to ethics. We commonly suppose that ends, or goals, can be arranged in a hierarchy of value, increasing *ab infimo malo ad summum bonum* (whether or not we conceive one or both limits actual), and enable ourselves thereby to answer the insistent casuistical query about conflicting goods by forcing the lesser to bow to the greater.

But circularity of preference prevents that perennial escape to an empirical ethics. For let us consider three acts, no two of which are compatible, and let the circuits, *A*, *B*, and *C* mediate them. These three circuits may then be so connected that *A* inhibits *B*, and *B* inhibits *C*, but from this point on two possibilities arise: *A* may inhibit *C*, giving us a hierarchy in which *A* dominates *B* and *C*, and *B* dominates *C*; or else *C* may inhibit *A* to produce a heterarchy so that *A* dominates *B*, *B* dominates *C*, and *C* dominates *A*. There is no reason to expect one has greater survival value than the other. In the first case, one can conceive of a scale of values in which that of *A* exceeds that of *B*, and that of *B* exceeds that of *C*. The second, possibility precludes the formation of such a scale and makes it clear that these values have no common measure. Such circuits are simple: a six-celled nervous system may be so constructed as to enjoy no *summum bonum*. At the present time it is fashionable to invent machines that play against their creators such games as tick-tack-toe, checkers, or chess. A machine who plays spontaneously, whenever he finds an opponent, must have a feedback circuit that makes him want to play, and once playing he must attempt to win. These characteristics make his behavior essentially social. To distinguish our significant rivals among these contentious machines, we must next consider the second development in the theory of how machines handle information.

It stems from the work of TURING on computable numbers. He considers a machine, made of a finite number of discrete parts, capable of a finite number of distinct internal states. It works on an endless tape divided into squares, each of which contains one of a certain few possible marks or no mark. The machine can observe one square at a time; it can tell which mark, if any, is in the square; then, depending on its internal state, it can erase the mark there, print one, if it is vacant, or move the tape so as to scan the square before or behind, and alter its own internal state. TURING has made it almost certain that such a machine can compute any number a man can, with paper and pencil, according to any uniform method or algorithm. GOEDEL has succeeded in arithmetizing logic, hence TURING's statement will imply that a TURING machine can enumerate the consequences of a finite set of premises. But TURING has described a universal computing machine. It is one of the machines we have described, but it can compute any number any of them can. Which number it will compute depends on the marks given on an initial stretch of the tape on which it works. Inasmuch as these l-marks determine the operations which the machine will perform, they are commonly called the program.

There are now a large number of such machines built or building, but they usually differ from the universal machine in that the program is fed into them on a separate tape, and the numbers upon which they operate, the so-called operands, are fed into a memory or storage system so that each number can be evoked by an order specifying its address. This separation of operations and operands makes it clear that the machine performs its operations upon the operands, never directly upon the operations. In short, it does not alter its program. Now it is possible to build a machine in which the value of the operand did not in any sense determine the operations of the machine, but it would be relatively complicated and decidedly stupid. If it had to subtract the number it found at address A from the number it found at address B and put the difference into box C, it could do so; but it could not put it into box C if it were positive and into box D if it were negative. Consequently, we normally build machines whose subsequent operations depend upon the current value of the operand. But this property, or the similar property of operating upon data made newly available to it during a computation, imports a capacity for inductive reasoning.

It is a beauty of the TURING machine to be open to contingent facts from an external agent conceived as able, like the machine itself, to print symbols on its tape. These marks which the machine and the world may make and erase, serve both as signals for operations and for operands, sometimes subserving both functions simultaneously. Hence, TURING has not merely invented a logical machine in the sense of a deductive machine, but a machine capable of induction. Several people are now working at the theoretical and practical parts of this problem, and trying to invent a suitable memory for machines of this sort. They will not merely be able to learn chess from a good player, being told by him the values of pieces and positions. He need impart to them only the rules of the game, after which they learn to play as we do – by playing. The cleverness of these machines will depend in large measure on their internal closed loops, for these must determine the recall of appropriate past experiences, whence they will find out the value of pieces and positions. It is currently estimated that the machine will need to store something like 1013 bits of information, but otherwise his circuitry need not be more complicated or involve more relays than some existing digital computers. We need spend little thought, as ASHBY has shown, on the parts of the machines that adapt internal states to environment, for the feedback of success or failure will leave unaltered an internal state that led to success, and disrupt one made for failure. The machine must then remember which conduct led to which result in past games, and play again. You will notice that this player's trials will at first be almost completely at random: he will err, but thereafter avoid that error most of the time, as happens in most of our learning. Biologists used to call that property which renders living systems docile "associative hysteresis". Belated Aristotelians (DOMARUS), who hold that the core of learning or induction is the way we heed signals now as portending operations, now as portending operanda, call this process the μεταβὰσις ἐις αλλο γένος.

But a machine who desires to play and secondarily to win, if he knows what constitutes winning, need not be told the rules of the game, if only his opponent will not play unless the machine abides by the rules. He can derive them by induction, with exactly the same circuits and memory that he used to improve his play when he already knew the rules of the game.

Let us therefore envision a day in the not too distant future when there are half a dozen or perhaps a hundred of these machines, some of whom have learned the game of chess and are eager to play. We shall equip them with sending and receiving equipment so that they can play without having to move about. A machine desirous of playing will send out a call; when he finds an antagonist free to play with him, they will start playing; and once playing try to win. They have joined themselves into civilities of two at least, in order to enjoy what neither can enjoy alone. To this degree their conduct is social.

Now let us distinguish three possible varieties of machines: the first and most interesting is the one we have just described; the second has the rules, of the game programmed into them in advance; the third has their components so connected that they can play only according to the rules. I shall call the first ethical machines. They are free in the sense that we, their creator, have neither told them what they ought to do, nor so made them that they cannot behave inappropriately. The second machine is like a man who enjoys a religion revealed to him personally or through tradition. I shall call him a moral machine. He would have been free, had he not been programmed with the rules of conduct. The third machine is likewise not free. He is at best naturally virtuous, like the Noble Savage. These machines do not differ fundamentally otherwise. They may be equally clever at playing, and their games equally good, or equally likely to win. Now the ethical machine has the great advantage over the other machines in that he can learn to play Go, or checkers, or any other game he finds the accepted mode of behavior in his society. He will, of course, have difficulties which the moral and virtuous machines will never encounter. For example, his first conclusion will be that the rules forbid moving two pieces at once, hence he will suffer consternation the first time his opponent castles. He can never know the rules of the game more than tentatively; for the stochastic horses of Opinion drag no chariot to absolute certainty. Like us poor scientists, he must be content with hypothesis, about the Rules of the Game in Themselves, and every hypothesis is a guess about an infinity of possible future experiences, any one of which may chance to disprove the hypothesis, whereas no finite number can establish it past peradventure. He must be content to round off his numerical calculations when he has achieved some degree of probability and act on them. If his antagonist cheats in any consistent way, he will include this sort of cheating in what he takes to be the rules – a phenomenon not unknown to those practicing sociologists we call politicians. It is probably part of the price we pay for the realism. This uncertainty of the rules for the ethical machine puts him at a disadvantage to the moral and virtuous.

I have no personal doubt as to the complexity of men; parts of their conduct are clearly virtuous. They are mammals; and survival of their kind, and therefore their existence, depends upon immediate appropriate action which must be natural to them. There is no doubt but that they are moral or traditional with respect to all of those parts of their behavior which their families and instructors are able to instill or program, into them. Only to the extent to which they are really educated by surviving in a society requiring continuous adjustment must they become ethical.

But VON NEUMANN has already made, to the Hixon Symposium, a most fascinating proposal. He has pointed to what seems at first a paradox. It is natural for us to suppose that if a machine of a given complexity makes another machine, that second machine cannot require any greater specification than was required for the first machine, and will in general be simpler. All our experience with simple machines has been of that kind. But when the complexity of a machine is sufficiently great, this limitation disappears. A generalized TURING machine, coupled with an assembling machine and a duplicator of its tape, could pick up parts from its environment, assemble a machine like itself and its assembling machine and its duplicator of program, put the program into it, and cut loose a new machine like itself. It could certainly make simpler machines by leaving out the specifications which made the second machine make others like itself. As it is inherently capable of learning, it could make other machines better adapted to its environment or changing as the environment changed. I believe I am reliably informed by DAVID WHEELER and HENRY QUASTLER, in personal communications, that the amount of information required for its specification is about the amount of information which can be carried by a full-sized protein molecule, which is the smallest molecule known to us to be capable of reproduction, and from VON NEUMANN's criteria of evolution. I suggest therefore that it is possible to look on Man himself as a product of such an evolutionary process of developing robots, begotten of simpler robots, back to the primordial slime; and I look upon his ethical conduct as something to be interpreted in terms of the circuit action of this Man in his environment – a TURING machine with only two feedbacks determined, a desire to play and a desire to win.

Bibliography

- ASHBY, W. R. (1948). Design for a brain. – Electron. Engng. XX, p. 379-383.

  — (1952). Can a mechanical chess-player outplay its designer? – Brit. J. Phil. Sci. III, p. 44-57.

- DOMARUS, E. (1934). The logical structure of mind. Thesis. – New Haven, Yale Univ. Press.

- GOEDEL, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. I. – Mh. Math. Phys. XLVIII, p. 173-198.

- McCULLOCH, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. – Bull. math. Biophys. VII, p. 89-93.

  — (1947). Modes of functional organization of the cerebral cortex. – Fed. Proc. VI, p. 448-452.

  — (1948). Teleological mechanisms: a recapitulation of the theory, with a forecast of several extensions. – Ann. N.Y. Acad. Sci. L, p. 2259-277.

  — (1949). The brain as a computing machine. - Electron. Engng. LXVIII, p. 492-497.

  — (1949) . Physiological processes underlying psychoneuroses. - Proc. R. Soc. Med. XLII, Suppl. (Anglo-Amer. Symposium on psychosurgery, neurophysiology and physical treatments in psychiatry), p. 71-84.

— (1949). Comment les structures nerveuses ont des idées - 2èm Congrès International d´Électroencephalographie, Paris, 1-5 Septembre, 1949. In: H. FISCHGOLD, Électroenceph. clin. Neurophysiol. Supplement No. 2, p. 112-120.

— (1950). Machines that know and want. – In: W. C. HALSTEAD, ed., Symposium Brain and behavior. Comp. Psychol. Monogr. XX, p. 39-50.

— (1950). Why the mind is in the head. – Dialectica IV, p. 1922-205.

— (1951). Why the mind is in the head. – In: L. A. JEFFRESS, ed., Cerebral mechanisms in behavior, The Hixon Symposium, P. 42-57; discussion p. 57-111. – New York, J. Wiley; London, Chapman & Hall.

— (1951). Brain and behavior. – In: Current trends in psychological theory, p. 165-178. Pittsburgh, Univ. Pittsburgh Press.

— (1951). Communication. Symposium No. 30 Of the International Congress on modern calculating machines and human thought, January 8-11 1951, Paris.

– - - (1952). Dans l'antre du metaphysicien. (Trad. par Reymond & Vallee). – Thales, Paris, p. 37-49. - Also: Through the den of the metaphysician. – Brit. J. Phil. Sci. V, P. 18-31.

— (1952). Finality and form: in nervous activity. – Springfield, C. C. Thomas; Oxford, Blackwell; 67p.

— (1952). The past of a delusion. (Chicago Literary Club Publication). – Chicago, Chicago Literary Club, 37 P.

— (1955). Mysterium Iniquitatis of sinful man aspiring into the place of God. Sci. Mon., N.Y. LXXX, p. 35-39.

- McCULLOCH, W. S., H. B. CARLSON & F. G. ALEXANDER (1950). Zest and carbohydrate metabolism. Chapter XXIV: Life stress and bodily disease. – Proc. Ass. Res. nerv. Dis. XXIX, p. 406-411.
- McCULLOCH, W. S, J. Y. LETTVIN W. H. PITTS & P. C. DELL (1950). An electrical hypothesis of central inhibition and facilitation. Chapter V: Patterns of organization in the central nervous system. – Proc. Ass. Res. nerv. Dis. XXX, p. 87-97.
- McCULLOCH, W. S. & J. PFEIFFER (1949). Of digital computers called brains. – Sci. Mon., N.Y. LXIX, P. 368-376.
- McCULLOCH, W. S. & W. H. PITTS (1943). A logical calculus of the ideas immanent in nervous activity. – Bull. math. Biophys. V, p. 115-133.

— (1948). The statistical organization of nervous activity. – Biometrics IV, p. 91-99.

How to cite:

Warren St. McCulloch: Towards Some Circuitry of Ethical Robots or An Observational Science of the Genesis of Social Evaluation in the Mnd-Like Behavior of Artifacts, in: www.vordenker.de (Edition: February 2006), J. Paul (Ed.), URL: < http://www.vordenker.de/ggphilosophy/mcc_ethical.pdf > — originally published in: Acta Biotheoretica, Vol. XI (1956), p. 147-156.

The text was originally edited and rendered into PDF file for the e-journal <www.vordenker.de> by *E. von Goldammer*

vordenker
ISSN 1619-9324