vordenker
Februar / März 2019

Toward Some Circuitry of Ethical Robots or an Observational Science of the Genesis of Social Evaluation in the Mind Like Behavior of Artifacts

- bilingual -

Zu Schaltkreisen ethischer Roboter oder: Eine Beobachtungswissenschaft der Genese sozialer Bewertungen im verstandesähnlichen Verhalten von Artefakten

Warren Sturgis McCulloch

How to cite:
Warren S. McCulloch, Toward Some Circuitry of Ethical Robots or an Observational Science of the Genesis of Social Evaluation in the Mind Like Behavior of Artifacts, Acta Biotheoretica, Vol XI, 1956, pp. 147-156,
reprinted in: Warren S. McCulloch, Embodiments of Mind, Cambridge, Mass., 1970, pp. 194-202
-
New German translation by Joachim Paul (with the help of deepl.com)

online: www.vordenker.de Neuss 2019, J. Paul (Ed.), ISSN 1619-9324
URL des Beitrags: < https://www.vordenker.de/ggphilosophy/mcc_ethical_en_ger.pdf >

vordenker
ISSN 1619-9324

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, 1953). 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

Zu Schaltkreisen ethischer Roboter oder: Eine Beobachtungswissenschaft der Genese sozialer Bewertungen im verstandesähnlichen Verhalten von Artefakten

by W. S. McCulloch

Vorlesung zur 13. Konferenz über Wissenschaft, Philosophie und Religion, New York, September 1952, und zur Sitzung unter der Schirmherrschaft des Department of Experimental Psychiatry, University of Birmingham, England, 1953.

Behauptung. Kein Empiriker erwartet, in Menschen oder Maschinen Ausnahmen von Naturgesetzen zu finden, aber die Naturwissenschaften sind nicht so konstruiert, dass sie die Probleme der Biologie, Psychologie oder Soziologie, die adaptives, perzeptives, aufmerksames oder kommunikatives Verhalten beinhalten, erklären oder lösen. In jüngster Zeit sind zwei neue Wissenschaften entstanden, in denen diese Probleme angesprochen werden können. Die Informationstheorie, die sich mit der Menge der Information beschäftigt, die von Signalen in der Gegenwart von Rauschen übertragen wird, unterscheidet zwischen Signalen, die wahr oder falsch sind, und Rauschen, das keine physikfremde Unterscheidung ist, sondern entscheidend für die Gestaltung von Kommunikationssystemen und anwendbar für alle Wandler von Informationen und alle Computer, seien es Menschen oder Maschinen. Die Kybernetik entstammt WIENER´s Einblick in Regler und Servosysteme: In der negativen Rückkopplung, durch die ihr Output ihren Input verringert, ist das, was zurückzugeben ist, keine physikalische Angelegenheit wie z.B. Energie, sondern lediglich Information über das Ergebnis einer früheren Aktion. Diese neue Wissenschaft hat eine Theorie des gesamten homöostatischen und zielgerichteten Verhaltens hervorgebracht, unabhängig von der physikalischen Natur der Komponenten, die die Wirkungsweise des Schaltkreises unterstützen, und ist daher sowohl auf den Menschen als auch auf Maschinen anwendbar. Hunderte von Artikeln belegen den Einsatz in der Biologie (STUMPERS, 1953). Dies ist ein bescheidener Versuch zu zeigen, dass diese beiden neuen Wissenschaften eine Bedeutung für soziale Probleme haben.

In Übereinstimmung mit dem englischen Sprachgebrauch bezeichnet das Wort "ethics" in diesem Artikel den Charakter oder die Art und Weise des Verhaltens, die sich im sozialen Verkehr entwickelt und den Zielen dient, die von dieser Gesellschaft geschaffen wurden. Mein Gesell-

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.

schaftsmodell ist das kleinste, das an einem solchen Ziel teilhaben kann, nämlich zwei Teilhaber. Drei, wie VON NEUMANN‘s Spieltheorie deutlich zeigt, erlauben es, dass sich je zwei gegen den Dritten verbünden können, und größere Mitgliederzahlen ergeben eine größere Komplexität.
Mit dreien in einer Demokratie – jeder bevorzugt sich selbst gegenüber dem Nachbarn zu seiner Rechten, und diesen gegenüber demjenigen zu seiner Linken – hängt derjenige, der gewählt wird, wenn nur zwei nominiert werden, ganz klar auf eine zirkuläre Art und Weise davon ab, welche zwei nominiert sind. Während wir bereits wissen, dass die Dynamik von drei und größeren Gruppen in kontrollierten Situationen davon abhängt, welche Kommunikationskanäle erlaubt sind, sind wir uns der Natur dieser Abhängigkeit noch nicht bewusst und können das Ergebnis nicht vorhersagen. Daher sollte das, was ich hier gesagt habe, nicht mit einer Analyse einer sozialen Situation verwechselt werden, an der mehr als zwei Personen beteiligt sind.

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.

Außerdem habe ich "moralisch" im vertrauten englischen Sinne für jene Verhaltensweisen verwendet, die der Anweisung in Gesetzen, oder der Offenbarung der Gesetze, des richtigen Verhaltens entsprechen. Nachdem wir also eine Fabel gehört haben, fragen wir: "Was ist die Moral dieser Geschichte?" Denn ich möchte sowohl die Entstehung von Werten aus bestimmten Erfahrungen als auch die Akzeptanz eines Wertesatzes von anderen rein instinktiven oder vererbten Wertesätzen unterscheiden, von denen angenommen wird, dass Insekten sie befolgen. Wenn der Leser für diese Unterscheidungen besser geeignete Begriffe findet, werde ich sie gerne verwenden.

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.

Aber in meinem einfachen Modell habe ich zu viel angenommen. Offensichtlich, angesichts des Wunsches zu spielen und eines Gegners, der aufgibt, wenn der Anfänger zu viele verlierende Züge macht, wird der Anfänger lernen, was es bedeutet zu gewinnen und wird zu diesem Zweck spielen.

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

Damit Sie nicht irregeführt werden, denken Sie bitte daran, dass die Wissenschaft zu Fragen von Gut und Böse nichts zu sagen hat. Aber unabhängig davon, ob der Mensch nun eine tautologische Theorie des Guten konzipieren kann oder nicht, wie Mathematik und Logik, ich meine eine normative Wissenschaft der Werte, kann er eine Beobachtungswissenschaft der Bewertung aufbauen. Um die Ursachen für ein bestimmtes Verhalten zu ermitteln, muss er die Entscheidungen der Organismen oder Maschinen beobachten. In diese Wahlen müssen, um ethisch einwandfrei zu sein, auch andere Organismen und Maschinen einbezogen werden, die Aufwand und

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.

Belohnung teilen müssen, sonst ergeben sich keine sozialen Fragen von Gut und Böse. Ich werde untersuchen, welche Maschinen durch Kooperation und Wettbewerb eine Gesellschaft bilden können, in der ihr Verhalten in einer Weise selbstdiszipliniert wird, die den Zielen ihrer Gemeinschaft dient. Zwei Entwicklungen in der Theorie der Maschinen und der Information werden uns helfen zu beginnen.

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.

Die erste von ihnen, obwohl alt in der Ingenieurskunst, ist neu in der Form, die WIENER, ROSENBLUETH und BIGELOW in ihrem Papier über teleologische Mechanismen vorschlugen. Es gibt viele mechanische und elektrische Geräte, die nach den von ihnen festgelegten Prinzipien arbeiten. In allen von ihnen erzeugt eine Änderung des Eingangs einen Ausgang, der auf den Eingang derart zurückwirkt, um diese Änderung zu vermindern. Alle geschlossenen Pfade dieser Art führen letztlich zu einem bestimmten Zustand des Systems. Sie bewirken, dass das Gerät immer dann in diesen Zustand zurückkehrt, wenn die Welt es aus diesem Zustand herausjagt. Sie werden deshalb als "fehlergesteuert" bezeichnet. Regler und Regulatoren sind in der Regel von diesem Typ. Zu denjenigen im Gehirn gehören einige, die ähnlich der automatischen Lautstärkeregelung eines Standard-Radioempfängers arbeiten. Bestimmte Reflexe, an denen sowohl das Gehirn als auch der Körper beteiligt sind, erfordern Schaltungen ähnlich denen in einem "Netzteil", das den variablen Wechselstrom aus der Leitung entnimmt und Gleichstrom mit der von Röhren benötigten konstanten Spannung liefert. Aber die Schaltkreise, die hier von zentraler Bedeutung sind, durchqueren den Organismus oder das Tier und seine Umwelt. Sie zeigen Appetenzverhalten, da sie von einem Ziel oder einer Zielvorgabe aus der Umwelt "fehlergesteuert" sind. Ohne sie hätten weder ein Mensch noch eine Maschine einen Zweck, der über seine/ihre innere Neuordnung hinausgeht. In Bezug auf solche Schaltkreise können wir, wenn sie andere Menschen oder andere sich appetent verhaltende Maschinen durchqueren, die Zwecke begreifen, die Ethik erzeugen.

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

Die bloße Tatsache, dass ihre Kollegen Appetenzverhalten zeigen, erfordert, dass die Maschine sie als Maschinen mit Appetenzverhalten behandelt, auch wenn sie sie nur für ihre eigenen Ziele nutzen will. Dies bleibt hinter dem kategorischen Imperativ zurück; kann sich aber dennoch als eine ausreichende Grundlage für eine Ethik des aufgeklärten Egoismus erweisen. Wir werden bald zur notwendigen Aufklärung und ihren mechanistischen Grundlagen zurückkehren, aber im Moment halten wir inne, um über das nachzudenken, was manchmal als "Wertanomalie" bezeichnet wird (McCULLOCH, 1945).

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.

Damit meinen wir, dass ein Tier oder eine Maschine, denen nacheinander die Wahl zwischen jeweils zwei von drei unvereinbaren Zielen, A, B, C, angeboten wird, manchmal A statt B, B statt C, C statt A wählt, und zwar konsequent. Ich selbst bin dem in der experimentellen Ästhetik begegnet, wenn ich drei Rechtecke, die in 2, 3 und 5 gleiche Rechtecke aufgeteilt sind, durch Paarvergleiche untersuche. Tierpsychologen haben herausgefunden, dass, sagen wir, hundert männliche Ratten, denen allen für einen bestimmten Zeitraum Nahrung und Sex vorenthalten wurde, Nahrung gegenüber Sex bevorzugen, Sex gegenüber der Vermeidung von Schock und Vermeidung von Schock gegenüber Nahrung. Dass dies geschieht, ist von theoretischer Bedeutung für die Ethik. Wir gehen gemeinhin davon aus, dass Enden oder Ziele in einer Hierarchie der Werte angeordnet werden können, aufsteigend *ab infimo malo ad summum bonum* (unabhängig davon, ob wir nun eine oder beide Grenzen tatsächlich wahrnehmen oder nicht), und es uns dadurch möglich ist, die hartnäckige kasuistische Frage nach widersprüchlichen Waren zu beantworten, indem wir das Geringere zwingen, sich dem Größeren zu beugen.

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*.

Aber die Zirkularität der Präferenz verhindert diese beständige Flucht zu einer empirischen Ethik. Denn betrachten wir drei Akte, von denen keine zwei kompatibel sind, und lassen wir die Schaltkreise..., A, B und C sie vermitteln. Diese drei Schaltkreise können dann so verbunden werden, dass der Kreis A den Kreis B und der Kreis B den Kreis C hemmt, aber ab hier ergeben sich zwei Möglichkeiten: A kann C hemmen, was uns eine Hierarchie gibt, in der A die Kreise B und C und B den Kreis C dominiert; oder C kann A hemmen, um eine Heterarchie zu erzeugen, so dass A den Kreis B und B den Kreis C sowie C den Kreis A dominiert. Es gibt keinen Grund zu der Annahme, dass das eine einen größeren Überlebenswert hat als das andere. Im ersten Fall kann man sich eine Werteskala vorstellen, bei der der Wert von A den von B und der von B den von C übersteigt. Die zweite Möglichkeit schließt die Bildung einer solchen Skala aus und macht deutlich, dass diese Werte kein gemeinsames Maß haben. Solche Schaltungen sind einfach: Ein sechszelliges Nervensystem kann so konstruiert sein, dass es sich keines *summum bonum* erfreut.

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

Derzeit ist es Mode, Maschinen zu erfinden, die gegen ihre Schöpfer spielen, wie z.B. Tick-Tack-Toe, Dame oder Schach. Eine Maschine, die spontan spielt, wenn sie einen Gegner findet, muss über einen Rückkopplungskreis verfügen, der sie zum Spielen bringt, und wenn sie einmal spielt, muss sie versuchen zu gewinnen. Diese Eigenschaften machen ihr Verhalten im Wesentlichen sozial. Um unsere bedeutenden Konkurrenten unter diesen um-

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

strittenen Maschinen zu unterscheiden, müssen wir als nächstes die zweite Entwicklung in der Theorie betrachten, wie Maschinen mit Informationen umgehen.

Sie stammt aus der Arbeit von TURING über berechenbare Zahlen. Er betrachtet eine Maschine, die aus einer endlichen Anzahl von diskreten Teilen besteht und zu einer endlichen Anzahl von verschiedenen inneren Zuständen fähig ist. Sie arbeitet auf einem Endlosband, das in Quadrate unterteilt ist, die jeweils eine von wenigen möglichen Markierungen oder keine Markierung enthalten. Die Maschine kann zu einem Zeitpunkt nur ein Quadrat beobachten; sie kann erkennen, welche Markierung sich gegebenenfalls im Quadrat befindet; dann kann sie je nach ihrem inneren Zustand die Markierung dort löschen, eine, wenn sie leer ist, drucken oder das Band bewegen, um das Quadrat vor oder hinter sich abzutasten und ihren eigenen inneren Zustand zu ändern. TURING hat nahezu sicher belegt, dass eine solche Maschine jede beliebige Zahl berechnen kann, die ein Mensch mit Papier und Bleistift nach einer einheitlichen Methode oder einem einheitlichen Algorithmus berechnen kann. GOEDEL ist es gelungen, die Logik zu arithmetisieren, daher impliziert TURING’s Aussage, dass eine TURING-Maschine die Folgen einer endlichen Menge von Voraussetzungen auflisten kann. Aber TURING hat eine universelle Rechenmaschine beschrieben. Es ist nur eine der Maschinen, die wir beschrieben haben, aber sie kann jede Zahl berechnen, die alle diese Maschinen berechnen können. Welche Zahl sie berechnen wird, hängt von den Markierungen ab, die auf einer ersten Strecke des Bandes, auf dem sie arbeitet, angegeben werden. Da diese l-Markierungen die Operationen bestimmen, die die Maschine ausführen wird, werden sie allgemein als Programm bezeichnet.

Es gibt heute eine große Anzahl solcher Maschinen, die gebaut oder gebaut werden, aber sie unterscheiden sich in der Regel von der Universalmaschine dadurch, dass das Programm auf einem separaten Band in sie eingespeist wird und die Nummern, auf denen sie arbeiten, die so genannten Operanden, in ein Speicher- oder Speichersystem eingespeist werden, so dass jede Nummer durch eine Bestellung unter Angabe ihrer Adresse hervorgeholt werden kann. Diese Trennung von Operationen und Operanden macht deutlich, dass die Maschine ihre Operationen auf die Operanden ausführt, niemals direkt auf die Operationen. Kurz gesagt, sie ändert ihr Programm nicht. Jetzt ist es möglich, eine Maschine zu bauen, bei der der Wert des Operanden in keiner Weise den Betrieb der Maschine bestimmt hat, aber es wäre relativ kompliziert und ausgesprochen dumm.

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 10^{13} 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

Wenn sie die Zahl, die sie bei Adresse A gefunden hat, von der Zahl, die sie bei Adresse B gefunden hat, abziehen und die Differenz in Feld C eintragen müsste, könnte sie das tun; aber sie könnte sie nicht in Feld C eintragen, wenn sie positiv ist, und in Feld D, wenn sie negativ ist. Daher bauen wir in der Regel Maschinen, deren nachfolgende Operationen vom aktuellen Wert des Operanden abhängen. Aber diese Eigenschaft oder die ähnliche Eigenschaft, mit Daten zu arbeiten, die ihr während einer Berechnung neu zur Verfügung gestellt werden, schließt ein Vermögen zu induktivem Schließen ein.

Es ist eine Schönheit der TURING-Maschine, offen für Einflüsse eines externen Agenten zu sein, der so wie die Maschine selbst in der Lage ist, Symbole auf ihr Band zu drucken. Diese Markierungen, die die Maschine wie die Außenwelt schreiben und löschen kann, dienen sowohl als Signale für Operationen als auch für Operanden und erfüllen gelegentlich beide Funktionen gleichzeitig. TURING hat daher nicht nur eine logische Maschine im Sinne einer deduktiven Maschine erfunden, sondern auch eine induktionsfähige Maschine. Mehrere Personen arbeiten nun an den theoretischen und praktischen Teilen dieses Problems und versuchen, einen geeigneten Speicher für solche Maschinen zu erfinden. Sie werden nicht nur in der Lage sein, Schach von einem guten Spieler zu lernen, indem sie von ihm die Werte der Figuren und Positionen erfahren. Er muss ihnen nur die Spielregeln vermitteln, danach lernen sie, so zu spielen, wie wir es beim Spielen tun. Die Cleverness dieser Maschinen wird in hohem Maße von ihren internen geschlossenen Kreisläufen abhängen, denn diese müssen die Erinnerung an entsprechende vergangene Erfahrungen bestimmen, woraufhin sie den Wert von Figuren und Positionen herausfinden werden. Es wird derzeit geschätzt, dass die Maschine etwa 10^{13} Bit Informationen speichern muss, aber ansonsten muss ihre Schaltung nicht komplizierter sein oder mehr Relais beinhalten als einige bestehende digitale Computer. Wir müssen, wie ASHBY gezeigt hat, wenig nachdenken über die Teile der Maschinen, die interne Zustände an die Umgebung anpassen, denn die Rückmeldung von Erfolg oder Misserfolg wird einen internen Zustand, der zum Erfolg geführt hat, unverändert lassen und einen für Misserfolg geschaffenen Zustand unterbrechen. Die Maschine muss sich dann merken, welches Verhalten zu welchem Ergebnis in früheren Spielen geführt hat, und wieder spielen. Sie werden feststellen, dass diese Versuche des Spielers zunächst fast völlig zufällig sein werden: Er wird sich irren, aber danach diesen Fehler die meiste Zeit vermeiden, wie es bei den meisten unserer Lernaktivitäten

most of our learning. Biologists used to call that property which renders living systems docile "associative hysteresis". Belated Aristotelians (DOMARUS, 1934), 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 μεταβάσις ἐις αλλο γένος.

der Fall ist. Biologen nannten das eine Eigenschaft, die lebende Systeme fügsam macht, "assoziative Hysterese". Späte Aristoteliker (DOMARUS, 1934), die behaupten, dass der Kern des Lernens oder der Induktion die Art und Weise ist, wie wir in Signalen einmal Zeichen von Operationen sehen, unmittelbar darauf aber Zeichen von Operanden, nennen diesen Prozess μετάβασις εἰς ἄλλο γένος [*].

[*] Wikipedia: wörtl. *Wechsel in eine andere Gattung* oder *Übergriff in ein anderes Gebiet.*

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.

Aber eine Maschine, die spielen und sekundär gewinnen will, wenn sie weiß, was den Sieg ausmacht, braucht nicht mit den Spielregeln vertraut gemacht zu werden, solange ihr Gegner nur dann spielt, wenn die Maschine sich an die Regeln hält. Sie kann sie durch Induktion ableiten, mit genau den gleichen Schaltkreisen und dem gleichen Speicher, mit denen sie ihr Spiel verbessert, wenn sie die Spielregeln schon kennt.
Stellen wir uns daher einen Tag in nicht allzu ferner Zukunft vor, an dem es ein halbes Dutzend oder vielleicht hundert dieser Maschinen gibt, von denen einige das Schachspiel gelernt haben und spielbereit sind. Wir werden sie mit Sende- und Empfangsgeräten ausstatten, damit sie spielen können, ohne sich bewegen zu müssen. Eine spielbegierige Maschine sendet einen Anruf aus; wenn sie einen Antagonisten findet, der mit ihr spielen kann, beginnen sie zu spielen; und wenn sie einmal gespielt haben, versuchen sie zu gewinnen. Sie haben sich zu einer Gesellschaft von mindestens zwei Personen zusammengeschlossen, um das zu genießen, was beide allein nicht genießen können. In diesem Maße ist ihr Verhalten sozial.

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

Unterscheiden wir nun drei mögliche Varianten von Maschinen: die erste und interessanteste ist die, die wir gerade beschrieben haben; die zweite hat die Regeln des Spiels, das im Voraus in sie programmiert wurde; die dritte hat ihre Komponenten so verbunden, dass sie nur nach den Regeln spielen können. Ich werde die ersten ethischen Maschinen nennen. Sie sind frei in dem Sinne, dass wir, ihre Schöpfer, ihnen weder gesagt haben, was sie tun sollen, noch sie dazu gebracht haben, dass sie sich nicht unangemessen verhalten können. Die zweite Maschine ist wie ein Mensch, der einer Religion anhängt, die ihm persönlich oder durch Tradition offenbart wurde. Ich werde sie eine moralische Maschine nennen. Sie wäre frei gewesen, wenn sie nicht mit den Verhaltensregeln programmiert worden wäre. Die dritte Maschine ist ebenfalls nicht frei. Sie ist bestenfalls von Natur aus tugendhaft, wie der edle Wilde. Diese Maschinen unterscheiden sich sonst nicht grundlegend. Sie können gleichermaßen

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.

clever im Spielen sein, und ihre Spiele sind gleichermaßen gut oder sie werden mit gleicher Wahrscheinlichkeit gewinnen. Jetzt hat die ethische Maschine den großen Vorteil gegenüber den anderen Maschinen, dass sie lernen kann, Go zu spielen, oder Dame, oder jedes andere Spiel, das zu den akzeptierten Verhaltensweisen in ihrer Gesellschaft gehört. Sie wird natürlich Schwierigkeiten haben, auf die die moralischen und tugendhaften Maschinen nie stoßen werden. Zum Beispiel wird ihre erste Schlussfolgerung sein, dass die Regeln verbieten, zwei Figuren auf einmal zu bewegen, so dass sie bestürzt sein wird, wenn der Gegner zum ersten Mal eine Rochade macht. Sie kann die Spielregeln nie mehr als versuchsweise kennen; denn die stochastischen Pferde der Überzeugung können keinen Streitwagen in absolute Sicherheit ziehen. Wie wir armen Wissenschaftler muss sie sich mit einer Hypothese über die Spielregeln an sich zufrieden geben, und jede Hypothese ist eine Vermutung über unendlich viele mögliche zukünftige Erfahrungen, von denen jede die Chance hat, die Hypothese zu widerlegen, während keine endliche Zahl je ihre Zufälligkeit beweisen kann. Sie muss sich damit begnügen, ihre numerischen Berechnungen abzurunden, wenn sie eine gewisse Wahrscheinlichkeit erreicht haben, und danach handeln. Wenn der Gegner konsequent betrügt, wird sie diese Art des Betrugs in die Regeln einbeziehen, die sie als ein Phänomen betrachtet, das jenen praktizierenden Soziologen, die wir Politiker nennen, nicht unbekannt ist. Es ist wahrscheinlich ein Teil des Preises, den wir für den Realismus zahlen. Diese Unsicherheit der Regeln für die ethische Maschine bringt sie in Nachteil gegenüber der Moralischen und der Tugendhaften.

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.

Ich habe keinen persönlichen Zweifel an der Komplexität der Menschen; Teile ihres Verhaltens sind eindeutig tugendhaft. Sie sind Säugetiere; und das Überleben ihrer Art und damit ihrer Existenz hängt von unmittelbarem und angemessenem Handeln ab, das für sie natürlich sein muss. Es besteht kein Zweifel, dass sie moralisch oder traditionell in Bezug auf all die Teile ihres Verhaltens sind, die ihre Familien und Ausbilder ihnen beibringen oder programmieren können. Nur in dem Maße, in dem sie wirklich gebildet werden, indem sie in einer Gesellschaft überleben, die eine kontinuierliche Anpassung erfordert, müssen sie ethisch werden.

But VON NEUMANN has already made, to the Hixon Symposium, a most fascinating proposal. He has pointed to what seems at first paradox. It is natural for us to suppose that if a machine of

Aber VON NEUMANN hat bereits zum Hixon-Symposium einen faszinierenden Vorschlag gemacht. Er hat auf das hingewiesen, was zunächst als ein Paradoxon erscheint. Es ist für uns selbstverständlich anzunehmen,

a given complexity makes another machine, that second machine cannot require any greater specification that 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.

dass, wenn eine Maschine mit einer bestimmten Komplexität eine andere Maschine herstellt, diese zweite Maschine keine größere Spezifikation erfordern kann, als für die erste Maschine erforderlich war, und im Allgemeinen einfacher sein wird. Alle unsere Erfahrungen mit einfachen Maschinen sind in diese Richtung gegangen. Aber wenn die Komplexität einer Maschine ausreichend groß ist, verschwindet diese Einschränkung. Eine generalisierte TURING-Maschine, gekoppelt mit einer Montagemaschine und einem Kopierer ihres Bandes, könnte Teile aus ihrer Umgebung aufnehmen, eine Maschine wie sie selbst und ihre Montagemaschine und ihren Kopierer des Programms zusammenstellen, das Programm in sie einsetzen und eine neue Maschine wie sie selbst freisetzen. Sie könnte sicherlich einfachere Maschinen produzieren, indem sie die Spezifikationen weglässt, die die zweite Maschine dazu befähigten, andere Maschinen wie sie selbst zu machen. Da sie inhärent lernfähig ist, könnte sie andere Maschinen herstellen, die besser an ihre Umwelt angepasst sind oder die sich mit der Veränderung ihrer Umwelt ändern können. Ich glaube, dass DAVID WHEELER und HENRY QUASTLER mich in persönlicher Kommunikation zuverlässig darüber informiert haben, dass die für ihre Spezifikation erforderliche Informationsmenge der Informationsmenge entspricht, die von einem großen Proteinmolekül, dem kleinsten uns als reproduktions-fähig bekannten Molekül, und von VON NEUMANN's Evolutionskriterien getragen werden kann. Ich behaupte daher, dass es möglich ist, den Menschen selbst als ein Produkt eines solchen evolutionären Prozesses der Entwicklung von Robotern, gezeugt von einfacheren Robotern, zurück bis zum Urschleim, zu betrachten, und ich sehe sein ethisches Verhalten als etwas, das in Bezug auf die Kreislaufaktion dieses Menschen in seiner Umwelt gesehen werden sollte – eine TURING-Maschine mit nur zwei bestimmten Rückkopplungen, einem Wunsch zu spielen und einem Wunsch zu gewinnen.

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 I'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; 67 p.

– (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.

vordenker
ISSN 1619-9324