POLANYIAN ARCHIVES

© POLANYIANA 2010/1-2. (19): 87–108.

Turing

We are interested in machines which test the description numbers (D.N.s) of other machines to see whther they are 'satisfactory', i.e. wh ether thex machines concerned print an infinity of 0s and 1s.

Let $\mathcal{M}$ be one such machine (for testing). Arrange the D.N.s which are guarenteed satisfactory by $\mathcal{M}$ in the order of the values of the D.N. or of the number of steps required for testing the machine, whichever may be the greater. In the groups within which the greater of these quantities has a constant value we may arrange according to the value of the smaller of the two. Let n_r be the r th number found satisfactory in this order. Consider the function $1 - \varphi_{n_r}(r)$. We can easily see how to produce a machine to calculate this function, giving the values in the form of a sequence of 0's and 1's whose r th element is $1 - \varphi_{n_r}(r)$. Let the D.N. of this machine be R. Clearly R is satisfactory. If $\mathcal{M}$ succeeds in finding out that R is satisfactory let $R = n_K$. The r th digit produced by the machine in question will then be $\varphi_{n_K}(r)$ as well as being $1 - \varphi_{n_r}(r)$. Hence for all r we have $\varphi_{n_K}(r) = 1 - \varphi_{n_r}(r)$ But now taking $r = K$ we have $1 = 2\varphi_{n_K}(K)$ which cannot be satisfied by putting either $\varphi_{n_K}(K) = 0$ or $\varphi_{n_K}(K) = 1$. Hence $\mathcal{M}$ cannot tell us that R is satisfactory.

In view of this we can design a machine $\mathcal{M}'$ which gives the same answer as $\mathcal{M}$ for all D.N.s except R but for R states that it is satisfactory.

MICHAEL POLANYI ON MIND AND MACHINE

George Polya to Michael Polanyi[1]

4.2.49

My dear Misi,

Many thanks for Your letter of 28.1.49. The paper of (A.M.) Turing you mean may be that in the Proceedings of the London Mathematical Society, v. 42, 1937, p. 230 "On computable numbers, with an application to the Entscheidungsproblem." It would cost me an effort much over my possibilities to understand this paper in details. Yet it gives me no reason to change my general attitude. He proposes to show "that there can be no machine which, supplied with any one R of these formulae [of the "functional calculus" of formal logic] will eventually say whether R is probable." [p. 259]. That is, although perhaps a machine may be built (a method devised) to decide whether *a given* purely logical statement is generally valid or not, still, once the machine *is* built, there are always some logical statements the validity of which cannot be decided by the machine (by purely mechanical application of the method). Therefore the machine you are writing about "which in due course will draw all the conclusions of any given system of axioms" will run on indefinitely and, *unless you add some new idea*, you *cannot* predict *how soon* it will prove or disprove a given statement. – There is room enough for heuristic!

"Les grands esprits se rencontrent." I was just about to write you, as your letter arrived, about the plans of my European trip. We are scheduled to arrive in Southampton on April 18 (Queen Elisabeth). I have a lecture in the London Math. Society scheduled on April 28, after which we shall leave immediately. It would be *very* important to me to *meet you* and *give one or two little talks* to some appropriate audience about "heuristic and inductive logic". I should like to have the reactions of some understanding people to my ideas – something which I cannot have here at all. I must visit Cambridge and stay in London, the time is short, and Stella, although she likes to travel, can stand very little of it without some kind of sea-, train- or other sickness. Therefore, I would much prefer some opportunity to talk in London or Cambridge – e.g. in the "seminar" of a friend of yours, if possible. If I have some time to plan my talk, I can adapt myself to pretty different audiences, I think. – I would like also to meet again Koestler and have a little more time to talk with him than I had in Palo Alto, provided that he is still interested in some version of "heuristic" – I mean he is a very amusing person and I would like to see him anyway, but I would not wish to take up his time if he has dropped his interest in the subject.

[1] Box 5, Folder 6.

From you, I should like to hear about everything, not only about "induction & heuristic", but also about persons, the general situation and – everything.

I hope, I shall see you.

As ever

P. Gyuri

Michael Polanyi to Max Newman[2]

Department of Economic and Social Studies

11th February 1949.

Professor Max Newman,
Department of Mathematics.

Dear Max,

I have talked to Miss Emmet and found her quite happy about a joint meeting of the Seminars of Philosophy and Mathematics under your aegis in the Department of Mathematics. I have also written to Polya, telling him that he will receive from you an invitation for the 27th of April. However, I shall *not send off this letter* until you have sent yours, and would like to ask you to ring through to Miss Olive Davies (915) when you have made up your mind finally and sent off your invitation. I shall be away from Tuesday to Friday during this week.

Polya's address is: Professor G. Polya, Stanford University, California.

Yours ever,

[Polanyi]

Michael Polanyi to George Polya[3]

Department of Economic and Social Studies,
The University, Manchester 13. ENGLAND.

11th February 1949.

[2] Box 5, Folder 6. With typed note: Returned with Professor Newman's compliments and thanks.

[3] Box 5, Folder 6. Addressed to Professor G. Polya, Stanford University, CALIFORNIA. With handwritten note: sent off 16th Feb. in agreement with Prof. Newman

My dear Gyuri,

I am grateful for your notes on Turing's paper. I have discussed the matter further with Newman and I think this clears it up completely.

It is wonderful to look forward to meeting you again in April. I feared I had missed you already. Max Newman will send you an invitation to address his Mathematical Seminar, sitting in conjunction with the Philosophy Seminar of Professor Emmet, on the afternoon of 27th of April. Our term starts on the 26th and an earlier date could not be arranged. I do very much hope that you will be able to manage to come. Newman will, of course, offer to pay your expenses, and you will stay with us at Hale.

I could try to arrange for you to address also the Society for Philosophy of Science, which has its headquarters in London, but even they may be hampered by the fact that the available period is mostly not in term time. However, I shall talk to Professor Dingle in London next week, and find out what he thinks about the matter. But I do hope that you will come to Manchester in any case for the meeting on the 27th of April.

Yours ever,

[Polanyi]

George Polya to Michael Polanyi[4]

26.2.49

My dear Misi,

Many Thanks for your letter. I am really very much obliged to you for all your trouble. I received Newman's letter in the same time as Yours, and Prof. Dingle's letter a few days later. These two opportunities to discuss my ideas will be a great help – I am much in your debt.

I hope that neither you nor Newman found it immodest that I asked him to move forward the talk one day, to the 26th of April, if possible. I thought it over: the *morning of the* 27th would be just as good, even better. I hate to be so particular (to ask for an "Extrawurst") but here is the matter: I am scheduled to talk Thursday the 28th in the London Mathematical Society. This is an "official" date and cannot be changed. I really should not spoil this talk, yet there is some danger of doing so, if I have to rush from the train to the talk or if I spend the foregoing night in the train (I have often pretty bad insomnia anyway). All what I wish is to arrange my visit to Manchester so that I can arrive in London in the evening of Wednesday, the 27th. I hope that this does not give too much trouble to you or Newman and will not appear as impudence.

[4] Box 5, Folder 6.

Stella would like very much to see both you and Magda, but she gets easily sea-, train-, and car-sick, and very badly too. (It is fortunate, that she never rides a camel.) She will leave the crossing of the Atlantic just behind her, that of the Channel just ahead, and will scarcely have the courage to make the trip to Manchester. I shall, however, be very grateful for your hospitality – I should like only to ask Magda to take a minimum of trouble.

With kindest regards from both of us to both of you

Yours ever

Gyuri

Michael Polanyi
Notes on the inexhaustibility of the mind[5]

In this brief summary I shall put down rather lossely my various points of departure, leaving open the question how far the argument derived from them could be improved (or may, on the contrary, show up an internal weakness) by a sustained analysis of the terms in which I have expressed it for the moment.

I take it that the scope of computing machines is identical with that of a formalised procedure, i.e. one which handles symbols according to strict rules and also announces results – such as 'A is a proof of B' – by the application of strict rules. I suggest that while no machine is certain to operate faultlessly, we may imagine one which does and that, while no rule is certain to be unambiguously applicable indefinitely, we may imagine that this is the case for some formal system that we have in mind. I suggest that the two propositions are equivalent.

(1) The programme of reducing mathematics to a specifiable set of axioms operated on according to specified rules has failed. The discoveries of Gödel (1930) have shown that arithmetic and advanced geometry are incomplete; for it has been possible to set up problems of an arithmetical or geometrical character that can be neither positively nor negatively decided in these systems. No extension of the axiom system can remedy this deficiency and no consistent and complete deductive theory, containing as its theorem all true sentences of arithmetic and advanced theory, can ever be constructed. Moreover, it is impossible to set up a general method which would allow us to differentiate between the sentences which can be proved and those which cannot be proved.[6]

In this passage reference is made to true mathematical statements which cannot be proved or disproved by specific rules of procedure from a given set of axioms. This refers to the fact that where such formal decision is impossible, we may yet feel

[5] Box 21, Folder 16. Date: 12th September 1949 It is an earlier version of the discussion paper (box 32 Folder6 dated from 13.09.1949.) that is discussed and published by Blum above (pp. 52-55.)

[6] From Tarski, Introduction to Logic, 2nd Edit. 1946, p.137-8.

compelled to accept a decision by other means and thus extend the pre-existing set of axioms. This again was discovered by Gödel. He showed that by reflecting on the operation by which a particular formula was proved to be undecidable within the pro-existing system of axioms, we are led to a compelling conclusion concerning the truth of that formula, which thereupon we will add to our axioms. This process can be extended indefinitely.

This establishes an inexhaustible programme for the discovery of ever more true mathematical formulae; by a procedure which is by its very nature incapable of formalisation, since at every new step it establishes as true a new formula which has been proved to be undecidable by the pre-given set of axioms and formal rules of operation. In other words, this programme can be carried out by the mathematician, but not by the calculating machine.

(2) The informal procedure by which we are capable of generating an indefinite number of new axioms which we believe to be true may be regarded as a process of reflection. From the aspect of our own mental operations we are led to new conclusions which lie outside the hitherto fixed range of these operations. This recalls that according to Poincaré[7] all mathematical innovation is essentially analogous to the procedure of 'mathematical induction'; and like this is based on "a bending back of the mind upon itself by which it observes its own mode of reasoning.

This immediately leads on to Tarski's work which shows that no self-consistent formal language can contain any semantic terms - like 'true', 'signifies', 'satisfies' etc. - which are applicable to sentences formed in that language. Which leads to the conclusion that in no given formalised system can the question be asked whether any statement made in terms of that system is 'true', 'significant' etc. (Though we may ask whether it is provable in that system). Hence the operation of any given self-consistent formal language precludes any process of reflection on the truth, significance, etc. of anything expressed in this language. Of such reflection we are in fact always capable and therein we manifest a mental power which lies beyond the operations of any pre-given formal system.

Tarski has shown that by exercise of this power we produce a new meta-language that is essentially richer than the one in which we made statements on which we are now reflecting in the meta-language. This essential enrichment is identical with the expansion of a pre-given system of axioms by the meta-mathematical process discovered by Gödel.

(3) I have made the assumption to start with that a system of definite axioms and strict rules of computation can exist and that this is equivalent to the existence of a machine which operates faultlessly. But the actual operations of the mind cannot be exhausted in terms of any set of pre-existent rules. For this I shall only quote a passage from Kant, as follows:

[7] Poincaré, "L'intuition et la logique en mathematique" (1900).

"If it were to attempt to show in general how anything should be arranged under these rules, and how we should determine whether something falls under them or not, this could only take place by means of a new rule. This, because it is a new rule, requires a new precept for the faculty of judgment, and we thus learn that, though the understanding is capable of being improved and instructed by means of rules, the faculty of judgment is a special talent which cannot be taught, but must be practised. This is what constitutes our so-called mother-wit, the absence of which cannot be remedied by any schooling. For although the teacher may offer, and as it were graft into a narrow understanding, plenty of rules borrowed from the experience of others, the faculty of using them rightly must belong to the pupil himself, and without that talent no precept that may be given is safe from abuse."

(4) It may be thought that the range of pre-given axioms and operations imparted to a machine might be enlarged by the interaction of the machine with its surroundings. We would have to exclude of course interactions with human beings observing the operations of the machine and giving it new instruction, based on their own reflections. The machine would have to learn something that is both new and true from its impact on inanimate nature or on people who know nothing about the calculations the machine is engaged in. From such experience it would have to adopt new rules of operation, which would in some sense be true.

I do not think that such a process could replace reflection as defined above, for it would be equivalent to a process of empirical induction. Viewed in this light, I suggest that the prospects of the machine are very limited. Induction cannot be formalized. All my beliefs as to what is established by experience contain an element of personal judgment, which as such cannot be explicitly stated. In all beliefs to be gained by new discoveries this element of personal judgment plays an even more conspicuous part.

All attempts at formalising induction break down through the fact that the result, if achieved, would be expressed in some fixed terms by which reference would be made to new experience. But no pre-given set of terms can claim such permanency. Major inductive discoveries (like relativity, quantummechanics, etc.) are made possible only by a fundamental change of the terms in which we refer to experience. This is in my view once more the kind of "essential self-enrichment" of which the reflecting mind is capable, but which is excluded by definition from the operations of any strictly formalised (or mechanised) logical system.

Max Newman to Michael Polanyi[8]

19 September 49

Dear Micha,

I am looking forward with much pleasure to the discussion on 20 October. From a letter to Jefferson of which Miss Emmet shewed me a copy it appears that you may be relying on her and me to do more in the way of arranging than either of us had realised. All we thought we had to do was to tell some members of our departments about it, and come ourselves. Is there anything more? I could fix a room if that would help. What time had you in mind?

I have read the notes, and it seems to me that your main conclusion is founded on a definite error. At the bottom of p.1 you assume that, since in the case of the Gödel example we are able to decide which of the two alternatives is right, we shall be able to do so in all the succeeding cases. There is no guarantee of this, and logic is stuck on a question of this very kind at present. In the Gödel case we know that "A is provable in the system" is false because it leads to a contradiction in the system, and by other complicated arguments (Gentzen 1936) which cannot be formulated in the system, we know that there is in fact no contradiction in the system. But a Gentzen may not always be forthcoming, and e.g. no answer is known at present, or in sight, to the question "Is the classical theory of real numbers consistent?" (The system for which Gentzen proved consistency is the theory of integers.)

In a more general way your argument seems to imply that some problems demonstrably unsolvable by machines are solvable by people, but I know of no instance of this, if "unsolvable problem" is used as in the Turing and Church theorems: indeed the discovery of any such instance would have very wide repercussions in logic. Finally there seems in your argument some tendency to identify the limitations of computing machines with the limitations of finite logical languages. Although the two have connections they are not at all the same: a machine is not tied to any particular logical language.

I will pass on the notes to Turing.

Yours sincerely

Max Newman

[8] Box 5, Folder 6. With Newman's handwritten note: It would be simplest if you sent *me* another copy when available.

Michael Polanyi to Karl Popper[9]

Dr. K. Popper,
The London School of Economics.

11th October 1949

My dear Karl,

I was delighted to receive your reprint. I have gone through your "Logik der Forschung" some time ago and shall be interested to see what principles you consider to be most central to its argument.

We are having a discussion in Manchester as between Mathematicians, Philosophers and Physiologists of the Brain, to which I would like to present the ideas outlined on the enclosed sheet. I should be grateful if you could spare a couple of hours to give me your opinion on this matter. If so I could go to London to meet you for this purpose. Since I have to be in Leicester on Monday, 24th October, it would be particularly advantageous to me if I could meet you in London on the night of the 25th or on the morning of the 26th. This may be a very impertinent demand, but I trust in your kindness to me that you will not resent it.

Hoping to see you soon, which is always a great joy to me,

Yours,

Michael

Michael Polanyi
Notes on Mind and Machine[10]

Since the conception of the machine involves some complications which I consider to be not strictly irrelevant to the argument (and I want to be as brief as possible) I shall limit myself essentially to the question whether the operations of a formalized deductive system might conceivably be considered equivalent to the operations of the mind. I believe that such a suggestion involves a logical fallacy.

The declared purpose of formalization is (1) to designate undefined terns, (3) specify unproved assorted sentences (axioms), and (3) strictly to prescribe the handling of assorted sentences which leads to the writing down of ne asserted sentences (proofs). There prevails throughout a desire to eliminate elements that are called 'psychological'. (1) 'Undefined terms' are chosen without aiming to signify commonly understood relations; (2) 'unproven asserted sentences' replaces 'statements believed to be self-evident'; and (3) the operations constituting 'formal proof' are intended to replace 'merely psychological' proofs.

[9] Box 5, Folder 6.

[10] Box 21, Folder 16. With typed note: Expanded from remarks made at the meeting on 27th October.

I think it is logically fallacious to speak of a *complete* elimination of what have been called 'psychological' but might better be called 'unformalised' elements of deductive systems.

(1) No undefined term can be introduced unless its use is first explained by ordinary speech or demonstrated by examples. It is a sign indicating the proper use which is to be made of it. The Acceptance of an undefined term implies, therefore, that we know its proper use though this is not to be formally described. This use is a skill of which we declare to be possessed.

(2) Similarly for 'unproved asserted sentences'. The fact that they are asserted is irrelevant. Asserted by whom? 'Asserted' merely disguises the unavowed yet indispensable 'believed': namely believed by the writer and held by him to be deserving universal belief. 'Sentence' is a disguise for statement: a statement about something. For if there is nothing that can satisfy a sentence there is no use deriving any other sentences from it; and only sentences that are statements can be satisfied or not satisfied. A 'proof' cannot be recognised as such unless it is true that whatever satisfies the axioms from which it starts will satisfy the theorems arrived at. In accepting a statement as an axiom we express the belief that we know what does and what does not satisfy it, and that everything does. Thereby we imply the unformalised knowledge of an indefinite range of operations and of their result.

(3) The mere handling of symbols according to the rules of formal proof constitutes a proof only to the extent to which we accredit these operations in advance with the power of carrying conviction. But 'proof (as Ryle would say) is a success-word. The success in this case lies in the capacity of the 'proof' to convince us (and to convince us also that others ought to share our convictions) that an implication has been demonstrated. No handling of symbols to which we refuse to award this success can be said to be a proof, no matter what pre-established rules it is said to conform to. And again the award of this success is a process which is not formalized.

Thus, I maintain, that a number of points, a formal system of symbols and operations functions as a deductive system only by virtue of informalised supplements. We must know the meaning of undefined terms, understand what is stated in our axioms and believe it to be true, and acknowledge an implication in the handling of symbols by formal proof. This knowing, understanding and acknowledging is not formalised and may be jointly designated as 'semantic operations' carried out in the 'semantic field' of the formalised system.

Formalization can be extended to hitherto unformalised semantic operations, but only if the resulting formal system can in its turn rely on yet unformalised semantic operations. The elimination of 'psychological elements' by formalization thus remains necessarily incomplete. The purpose of formalization lies in the reduction of informal functions to what we believe to be more limited and obvious operations; but it must not aim at their elimination.

The semantic operations attached to a formal system are functions of the mind which understands and correctly operates the system. To believe that I understand

and correctly operate a formal system implies that I know how to operate its unformalised functions. Since a formal system will always require supplementation by unformalised operations, it follows that none can ever function without a person who performs these operations. A formalised deductive system is an instrument which requires for its logical completion a mind using the instrument in a manner not fully determined by the instrument; while the mind of the person using the instrument requires no such logical completion. A person can carry out computations by the aid of a machine (or formal system) or without it, but a computing machine cannot be said to operate except within a system:[11]

I		II		III
mind	→	machine	→	things to which the machine informally refers

If in this system we replace 'machine' by another mind we have,

I'		II'		III'
mind (1)	→	mind (2)	→	things to which mind (2) informally refers

where the functions III' are those of mind(1), while mind(2) merely functions as an instrument of mind(l). This is of course the strictly behaviourist model of mind(3) which attributes to the observed mind an entirely different character from that required for its observation by mind(1).

[11] The shorter version of this text (see also Box 21, Folder 16) from this point follows as below:

I		II		III
mind	→	computing machine (or formalised system)	→	things to which the machine informally refers

I shall call this the system „F".

2. Semantic amplification of formal expressions.

The term II in the system F may be called a "detached" expression. The aim of the previous section was show that this detachment is only apparent and to indicate the context within which can alone sustain a formal expression. I have written down the system F to designate this context. In this designation the detached expression is amplified by the terms I and III flanking it on either side. The two are closely related, as the first refers to the mind and the second to an operation of the mind within the semantic field of II. I shall now try to link up this designation of the system F with earlier attempts at a formalized semantic amplification of detached formal expressions.

This inequality which arises from different temporary functions of the mind can be accepted only as representing two aspects of the mind. Both are significant and both are incomplete. The experimentally observed aspect of the mind ('mind(2)') is the brain surgeon's aspect of it. It excludes the unformalised functions of the mind and therefore lacks the feature of personal responsibility. Such responsibility is exercised by the observing mind ('mind(1)'), through its unformalised functions.

Only 'observing minds' (minds(l)) can be supposed to communicate with each other. Inter-personal dealings like listening to or addressing a person exclude the observing of one person's operations (mind(2)) by the other in the sense in which mind(l) experimentally observed mind(2).

Michael Polanyi
Clues Towards an Understanding of Mind and Body[12]

I have written on various occasions about the stratification of the universe, and have analyzed both the logical structure of our knowledge concerning consecutive levels and the way in which the workings of two such levels are related. By way of introduction I shall sum up the results of this enquiry in one paragraph, hoping that what follows later will make its meaning clearer.

Each level of existence is a comprehensive entity rooted in the level below it. When we focus our attention on a higher level we rely on our awareness of its particulars (forming a lower level) on which we are not focusing at the time. The higher level relies for its workings on the laws governing the particulars of the lower level and these laws also limit the range of the operations on the higher level and account for its failures. But the operations of the higher level cannot be accounted for in terms of the laws which govern the lower level, and hence the examination of the particulars of the lower level does not reveal the laws of the upper level. The higher level is unaccountable in terms of its particulars.

The extent to which the particulars of a lower level are specifiable and to which their connection forming a comprehensive entity can be explicitly described, varies from one type of system to another. A machine can be taken to pieces and each of its parts be examined in itself. This will not tell us how they combine to make the machine work, but the working relation of the parts can also be explicitly stated in terms of engineering.

Specifiability is much more limited in other cases. When we recognize the physionomy of a person and can say that it expresses fear, anger, boredom, or alertness, or when we make a difficult diagnosis, or exercise expert knowledge in

[12] Box 42, Folder 5. This text appeared in Good, I.J., ed. (with the help of A.J. Mayne and John Maynard Smith) *The Scientists Speculates: An Anthology of Partly-Baked Ideas*, London, Heinmann, 1962 (Basic Books, New York, 1963; Paperback, Capricorn Books, New York, 1965.)

identifying an unusual specimen of some kind, we are relying on our capacity to recognize a comprehensive appearance without being able to say what exactly are its particulars and how they combine to a comprehensive entity. This is true also for comprehensive entities of a practical kind, achieved by the exercise of a skill. We may swim, ride a bicycle, play the piano or the violin, without being able to identify some of the most important elements which constitute these actions, nor be able to tell what rules we are following in effectively combining these elements. And diagnosing and testing go together; skilful knowing and doing are always combined.

Some of the particulars of a lower level, and in certain cases all of them, may be observed in themselves, but thus observed, they convey no comprehension of the upper level. On the other hand, when the particulars on the lower level are noticed as clues to the comprehensive entity on the upper level they are not observed in themselves. These two kinds of awareness are mutually exclusive: we can either observe a particular uncomprehendingly, or else read it as a clue to a higher entity which it signifies to us.[13]

The higher animals form a particular class of comprehensive entities: they are distinct individuals governed by active centres. Embryological development is controlled by centres of growth; the vegetative functions on which life depends are controlled by the autonomous nervous system; the central nervous system has motoric and sensory centres and contains, less localisably, centres of intelligent behaviour. Finally, on the highest level, we recognize the human person, as the centre of responsible judgment.

Thus the particulars of the centrally controlled individual are the workings of its centres. This makes for a two-way relationship between these particulars, and the comprehensive entities formed by them. The animal appears constituted by the integration of particulars our awareness of which conveys to us our understanding of the animal as a centrally controlled individual. We observe a person's mind by reading the workings of his mind.

I have said that we can never be sure of identifying the particulars of a physionomy. I may add now that they could never be identified at all except by previously attending to the physionomy as a whole. Since the workings of the mind form such a physionomy, the behaviourist programme of attending to the particulars controlled by the mind, instead of to the mind, is impossible and indeed absurd. Ryle's conception of the mind, which identifies it with its workings, is equally unacceptable, since it fails to distinguish between the observation of these workings and the reading of them, though these two are mutually exclusive.

Growth and physiological functions are totally or predominantly unconscious, but we usually are aware of our own motoric and intellectual efforts. Yet we do not normally observe our body or our mind while we are using them. Instead, we

[13] Denotation represents a form of such signification, but this plays no part in the following argument.

are aware of our own exertions of mind and body in terms of the purposes we are pursuing or the objects to which we are attending. Thus our awareness of ourselves in action is related to our objectives, in the same way as our awareness of the parts of a comprehensive entity is related to our attention fixed on that entity. Our awareness of our own limbs in action is of the same subsidiary kind as our awareness of another animal's limbs in action; so that, as we live in our limbs in using them, so we live in another animal's limbs in observing his behaviour. This is also how we observe another person's mind. We read the acts of his mind by relying on our awareness of them in the same way as we are aware of our own mental efforts in pursuing our own objectives. All observation of life and mind is convivial.

Two major classes of human artefacts are also known by a manner of indwelling. We observe the principle of a machine by viewing its parts as organs - which is also the way its inventor first imagined the machine. We know a language not by attending to its sounds, but to its words. Indeed the stratification of speech goes further; we know sentences by attending subsidiarily to the words composing them, and we grasp the meaning of a discourse by attending subsidiarily to its sentences. This is how we know a work of art; its purpose is merely to be dwelt in, as in a vivid and intelligible extension of our being.

The fact that we know another mind by indwelling, means that we understand it as the agent of the same kind of understanding by which we understand it. The sight of a man's alert eyes and face instantly conveys to us the presence of a conscious, sane and intelligent mind, having the same faculties that we ourselves exercise as conscious, sane and intelligent beings. Principal among these faculties is the capacity to comprehend particulars in terms of coherent entities. Such intelligent combination of two kinds of awareness is an essentially conscious act. The effort to understand something and the subsequent achievement of seeing an aggregate of particulars as parts of a comprehensive entity can be ascribed only to a sentient being; no insentient automaton, however closely it may reproduce the signs of such a performance, can be said to strain its attention and to see certain things one way rather than another. The representation of man as an insentient automaton is indeed as contrary to our observation of other minds as it is to our experience of our own mind.

Remember now Laplace's vision of universal knowledge. He said that if at any moment we knew the positions and velocities of all particles of matter, and the forces acting between them, we could compute the positions and velocities of the same particles at any future or past moment, and thus all things to come and all things gone by would be revealed to us. This mechanical conception of the universe would have to be transposed today into quantum mechanical terms, but it would still recognize only one single level of existence, acknowledging no comprehensive entities nor the ensuing stratification of existence. This raises questions which I shall exemplify by the case of machines. The principles according to which a machine works, cannot be accounted for in terms of physics and chemistry. Yet the machine is an inanimate body. How can it be then that physics and chemistry

should fail to describe it fully? And if there do exist superior principles which control its comprehensive actions, how can these fail to interfere with the laws of physics and chemistry which apply to the parts? How can the machine actually rely on the laws of physics and chemistry for performing its functions as a machine?

The answer is that the laws of physics and chemistry do not determine the configuration of positions and velocities in which they start to operate. Laplace himself says that the initial conditions have to be given before the physicist can make any predictions. So any mechanical system can be shaped initially according to principles which are not accounted for by physics and chemistry, and it may then continue to function in accordance with these same principles while relying on the laws of physics and chemistry. A machine comprises two levels of existence because its initial parameters are controlled by the laws of technology which cannot be accounted for by the laws of physics and chemistry.

Insofar as the living body functions as a machine, these conclusions can be readily applied to it too. Physiology consists of operational principles relying on the laws of physics and chemistry which control the parts in which these principles are embodied. Physiology can therefore not be accounted for by the laws of physics and chemistry, any more than the operational principles of a machine can be. The operational principles of living beings are embodied in the parameters left indetermined by physics and chemistry – in the same way as in machines.

It is customary to identify the mechanistic explanation of living beings with an explanation in terms of physics and chemistry. We see now that this is mistaken. We should recognize that a living being, even when represented as a machine, comprises two levels of existence, of which the higher relies on the lower, without interfering with the laws governing the latter.

The only way, therefore, to reconcile a mechanistic conception of the universe with the fact that it has given rise to the evolution of living beings and eventually to the emergence of machines, is to assume that these comprehensive entities were preformed by a suitable pattern of parameters within the mass of primordial incandescent gases. Instead of rushing about at random, its particles must have been ordered by such a pattern of positions and velocities as would manifest itself, as the gas cooled down, by producing living beings and the whole evolutionary development, including man and all the works of humanity. Nothing would then be new; in the atomic structure of the primordial gases, the works of Shakespeare could have been legible to a Laplacean mind, provided it understood English.

But the assumption of such an infinitely sophisticated original gas would save the comprehensiveness of mechanics only by abandoning the randomness of thermal motions on which thermodynamics is based. And even so it would be useless. It could explain machines and living beings working as machines, but no ordered pattern of primeval gases could account for the sentience of living beings, since physics and chemistry know nothing of sentience in matter.

Being thus forced to abandon mechanical preformation, we must look out for some

acceptable conception of cosmic epigenesis. We find some clues of this in the fact that living beings cannot be represented altogether as machines. This seems true even on the vegetative level; embryological development can repair early mutilations by using material of the embryo resourcefully for the achievement of normal shapes and organs. Next, mutilated animals can immediately adapt themselves by producing suitable patterns of behaviour. These feats have been likened by gestalt psychologists to the way animals and men mentally reorganize the field of experience towards a new purpose. And such primitive integrations contain already the germs of radical intellectual innovations achieved by human genius.

Attempts have not been lacking to represent these innovative functions by machines, but these would at best lead back to a mechanical predeterminism that would ascribe to the primeval incandescent gases of the world a pattern so intricate that in it would be foretold the evolution of all living beings, and all the works of human genius, to the end of time - while it would also represent human beings once more as insentient automata.

The absurdities of mechanical predetermination appear so linked to the absurdity of representing human beings as robots, that the two must be eliminated simultaneously. Consciousness must be recognized as endowed with the capacity of acting as a first cause. At every stage of organic evolution at which there emerges a higher level of existence that cannot be logically accounted for by the laws governing the level below it, we must postulate the presence of an essential innovation initiated by a first cause, which, for the sake of continuity, we should endow with some degree of sentience.

Great intellectual innovations are more rapid than those taking place at the lower levels of evolution and they also have the greatest intensity of conscious effort. We may adjoin therefore to the hierarchy of existential levels an increasing intensity of innovations occurring between levels, and a growing consciousness of these innovations.

We observe the centre of a responsible judgment by entering into the thoughts of a person delivering it; we gain a glimpse of great minds by immersing ourselves in their masterpieces. The first causes observed here illustrate the fact that even at the highest level the scope of such causes is restricted. It is conditioned by parochial circumstances of time and place, and its tasks are limited to the exploitation of hidden possibilities of innovations. We may assume then – by continuity – that the task of innovating causes at lower levels is likewise restricted by the range of possibilities. Judging by analogy, we may conceive indeed that these causes are evoked by the proximity of yet unrealised potential levels of higher integration, just as our awareness of hidden knowledge presents our mind with problems and evokes our efforts to solve them.

One part or other of these conclusions has been anticipated by Butler, Lloyd Morgan, Bergson or Whitehead. Yet I feel that the logical distinction between existential levels lends a new coherence and compulsive force to the argument.

Michael Polanyi
Notes about Mind and Body[14]

(Following a discussion with Professor John MacCarthy of Stanford Pete Uttley and Meyer Schapiro on February 19, 1963)

1) Think of any performance of the human mind, and call it X. Can the machine do X? My answer is, that to the extent to which X defines a human performance, a machine is always conceivable which will do X or appear to do X.

2) Suppose X is a performance that we usually attribute to the working of the mind, i.e. an intelligent performance. Does it follow then that machines can think? No, I have admitted only that they might appear to think. But can we distinguish between 'thinking' and 'appearing to think', even though we in fact cannot distinguish the performance of a machine from that of a human being? Turing argued that we must identify the performances of a machine and a human being if, when put to a suitable test (by an experimenter who knows which is the machine and which the human person), we prove unable to tell the one from the other. This conclusion seems to follow from the theory of meaning which defines the meaning of a term as the observable manifestations of that which is meant by it. I shall call this the positivist theory of meaning and argue that its application is ambiguous in respect of the present case and is unsatisfactory in general.

3) As applied to the Turing experiment, the positivist theory of meaning is ambiguous. For it produces a different result for the experimenter than for the observer whom the experimenter is testing. The experimenter must know which is the human being and which the machine, and he may attribute powers of thought to the human being, even though the observer outside cannot tell its performance from that of the machine. He will do that, if he believes that there is more to the process of thought than can be revealed by any set of explicit performances. He may say that he reads in a man's eyes and features the presence of human thought which he cannot read in a machine which lacks expressive eyes and features. It is not clear, whether this would be contrary or not to the positivist theory of meaning.

4) Let me show then how this ambiguity of the positivist theory of meaning extends to the point where the theory breaks down. A malingerer complaining of severe headaches may deceive a doctor. Suppose that, up to a certain day, he has deceived all doctors; is there a final day on which he will have qualified as a genuine sufferer? Will he then be entitled to sympathy and help as if he were actually suffering? To deny this is to contradict the positivist theory of meaning, unless we include among 'observable manifestations' some manifestations which might not be observable by any specific future date. I shall say, therefore, that the distinction between a successful malingerer and a true sufferer remains valid *at all times, in*

[14] Box 6, Folder 3; Box 21, Folder 17.

view of the indeterminate expectations which we ascribe to the true sufferer as distinct from the malingerer.

5) If we define reality as that which, being real, might yet reveal itself in an indeterminate range of future manifestations, we may conclude that pain is real and that its consciousness is a real quality of conscious thought, which might be recognised as such, even though we may be unable to specify any particular performance by which its presence is identifiable. At this point the argument coincides partly with earlier statements (e.g. in all three Terry Lectures) about the reality of the mind. The coincidence is incomplete, for the earlier statements did not always explicitly refer to a conscious mind.

6) But it may be thought that a state of consciousness, which might not be identifiable by specific overt behaviour, would yet be accompanied by characteristic bodily manifestations inside the body. Could we not represent these physiological processes in terms of a mechanism, and thus give a complete account of mental processes in terms of a physiological machinery, the operation of which could, in principle, be duplicated by a computer?

I answer to this as follow: The neurophysiology of man tells us about his consciousness and sentience by acknowledging the authenticity of conscious and sentient states of mind previous to relating them to underlying bodily processes. A description of these neurophysiological processes without their bearing on the corresponding states of consciousness is a meaningless fragment of neurophysiology. A neurophysiological mechanism can be established only by bearing on these conscious processes, and it cannot replace that which it bears on, any more than the word 'table' can be used as a table. Furthermore, the word 'table' is meaningless if tables do not exist irrespective of being designated as such.

7) But we may ask: If conscious processes are the meaning of the neurophysiological mechanism that bears on them, why can we not *read* that meaning in the way we read pain in the features of a sufferer? I do not deny that we might develop a capacity for an empathic interpretation of some neurophysiological processes. Suffice to point out that this will not always be the case. We may have a complete knowledge of the neurophysiology of vision, yet we cannot expect that by attending to these processes in the sensory system of a seeing subject, we shall see what he sees by his relying on his own awareness of these processes. I can only repeat here that this fact „might be regarded as an instance of the destruction of meaning which takes place when we focus our attention on the isolated particulars bearing on a comprehensive entity... The subject's awareness of his own neural processes has a much higher grade of indwelling than the physiological observation of them." Hence he can see by them (or *from* them), while the physiologist can only look <u>at</u> them.

8) What then, should we think of an „homme machine", a representation of man as a machine? It means either (l) that we identify a person whose sentience we physionomically recognise, and share, with an automaton which accounts for any specifiable performances of the person, but is known to be a mere imitation of the

person; or (2) that we identify the sentient person with a complete physiological mechanism, the meaning of which consists in its bearing on the person's sentient behaviour, which we are trying to replace by it. We might then either choose to treat a person as if he were insentient, which would be wrong (and also inconsistent with the fact that our construction of the mechanism assumes an identification of our own sentient experiences with the stimuli forming part of this mechanism); or else choose to treat him as sentient, which would contradict our identification of him with a mechanism in which neural currents replace the sensations giving rise to them.

9) It is often said that to speak of sentience or of the corresponding neural mechanism, is to speak of the same thing in different languages. Actually, it is to speak of two different things, the relation of which is similar to that between a comprehensive entity and the particulars which constitute it. We will identify such an entity with its comprehensive features, but not with its particulars, which can be referred to without contradicting the existence of the comprehensive entity only as bearing on that entity.

There are cases in physics, e.g., Fermat's law, defining the path of a ray of light as the fastest of all possible neighbouring paths, where there is an equivalence of two very different descriptions of the same phenomenon. In such cases one can transform a statement from one language into another. Fermat's theorem can be shown to be equivalent to the differential equation determining the curvature of a ray of light as a function of the gradient of the refractory index at that point. No similar transformation is possible between the two languages referring respectively to a sentient person and his physiological mechanism.

10) The conclusion reached in these notes will yet have to be restated in a more systematic way. They ultimately rest on such conceptual assertions as: "An undetected malingerer does not become a genuine sufferer, however numerous the tests he satisfies." This fiduciary commitment must yet be more fully acknowledge.

11) The authentication of consciousness rests on the unspecifiability of tacit knowing. The two terms of tacit knowing are known in different ways; we are aware of the first in its bearing on the second. We are attending from the first to the second. This relation is defined by the intentional and attentional act which constitutes it. To acknowledge an intention or attention is to authenticate a conscious effort in using its powers of skilful integration. The effort is directed from the proximal to the distal, radially from the center of knowing. This integrative force projecting outward is intrinsic to tacit knowing for it is by this outward striving and power that tacit knowing is achieved and held to be true. By contrast, explicit inference proceeds in one single plane. Its premises and conclusions are placed equally on this plane, and the operations leading from the one to the other lie also in this plane: insofar as an inference is explicit it is also detached, impersonal, and can be operated mechanically.

Acts of tacit knowing are unspecifiable in the same sense that thinking differs from the operations of a computer, to which we can ascribe no efforts of intention or attention. Tacit knowing is a mental act which can be successfully replaced and imitated by an explicit procedure, but which can never be identified with such a procedure.

12) Here is another variant of my argument. Take any experiment of perceptual qualities, such as constancy of colour, or a perception resulting from a pair of differently coloured retinal images. Suppose we find a mechanism which will account for what we perceive. Represent then the organism as this neurophysiological machinery. What does the machinery perform? It produces a response to a set of stimuli. What is this response? We may identify it with a statement describing the perception in question. But the statement can be acknowledged as such only if it has a meaning and it has no meaning unless by referring to perception as a conscious experience.

The replacement of a mental process by a neurophysiological mechanism appears *plausible*, when the process can be defined with *fair approximation* by a behavioral performance (like learning a response to a sign). The *inherent inadequacy* of the replacement becomes evident when it refers to a mode of perception, and it becomes *manifestly absurd* when applied to the formation of an optical illusion. A machine which produces an optical illusion is a logical absurdity, unless regarded as a mechanism by which a particular state of consciousness is brought about in a living organism.

John McCarthy to Michael Polanyi[1]

Professor Michael Polanyi
Institute for Advanced Study in the Behavioral Sciences
Stanford, California

April 1, 1963

Dear Professor Polanyi:

Thanks for your notes on mind-body. Like you I do not accept a positivist theory of meaning. Nevertheless, I think it may be possible to understand intelligence, not just neuro-physiology well enough to duplicate it. I think that a successful thinking machine will have to have a hierarchical structure which will involve distinctions between "seeing" spots of color and seeing the dog they represent. In particular, I don't think that a machine that suffers from optical illusions is an absurdity. Namely, the

[1] Box 6, Folder 3. cc: Dr. A. M. Uttley, Dr. Meyer Schapiro

human mind (?), eye(?) has a great capacity to see an object of a familiar type while "seeing" only part of its surface and this under varied conditions of lighting, angle, and obscuration. A successful mechanism for going from "seeing" to seeing must jump to conclusions, sometimes incorrectly. The mechanism must also be capable of greater care when there is more time and more opportunity to "see" from different angles and so will sometimes recognize that it has fallen for an optical illusion.

I am sorry not to go into more detail, but I had better say something now rather than put it off till I have more time.

Best regards,

John McCarthy