Gotthard Günther [*]

Can Mechanical Brains Have Consciousness ?

If by mechanical brains you mean the modern calculators like Vannevar Bush´s Differential Analyser (which you could watch in action, in the film "'When Worlds – Collide") or one of the digital computers like ENIAC EDVAC, UNIVAC and MANIAC, let me assure you that *none* of them can think. Nor will any of the more advanced models which man may build along these lines in the next centuries. And the same holds for the recent designs of logical computers: neither they nor their technically most advanced descendants will ever think.

How do I know? I admit that you will, find a single cyberneticist or designer of a computing machine who would be prepared to make such a sweeping statement. The general attitude among scientists concerned can be described as follows: thinking is a specific form of consciousness; however, nobody knows what consciousness really is, let alone how it can be produced by mechanical means; so your guess is as good as mine. And, indeed, a lot of wild guessing is going on. Some people passionately maintain that mechanical contrivances will never acquire consciousness – ergo they won't think even after, Doomsday. Whereas the opposite party blithely advertises that robot brains of present design are on their way toward learning how to think, and that all is just a matter of time and a bit of patience.

Both opinions are – misguided by entirely unwarranted assumptions. The skeptical viewpoint of the first party implies that we are not only at the present stage of science completely ignorant of what consciousness really is, but that we shall never leave that state of ignorance – the reason being that consciousness is a manifestation of a metaphysical soul of man and therefore of divine origin. Of course, you can't design that. You might as well start with the blueprints for the archangel Gabriel !

The other group, however, assumes, equally erroneously, that we do not have to know what consciousness is: that it is just a word or label for the abstract sum of all our perceptional and apperceptional functions. Ergo, if we reduplicate all those functions of sensitivity, memory, learning, capacity to make decisions, quantitative and qualitative reasoning, etc., through the medium of mechanical procedures, we have produced consciousness and thinking in a man-made, machine .... because consciousness has no independent "physical" reality outside its own functions. Consciousness is as the nominalists say, a mere name by virtue of which we comprehend and lump together an extremly diversified array of brain actions under a general and abstract heading. There exist horses, dogs, birds and snakes; but there exists no animal. "Animal" is just name, and so is "consciousness."

This theory is equally false, since it has been discovered that consciousness is an existing and most intricate mechanism apart and separated from its own functional proceedings. This discovery is made and expounded in Kant's famous work "The

---

* Originally published in: Startling Stories, Vol.29, no.1, New York, 1953, p.110-116.

Critique of Pure reason," and from there a new scientific discipline has evolved, usually called: "transcendental logic."

The first systematic treatise of this new type of logic was Hegel's "Phenomenology of the Mind." But don't try to read it. It has been called the most difficult book ever written in the history of mankind. The English philosopher Hutschinson Stirling, who wrote a comprehensive book on Hegel and his logical theories, titled his work: "The Secret of Hegel," in consideration of the difficulties of the new type of logic. After Stirling's book was published the joke went around among logicians: if Hegel had a secret, then Stirling kept it well. Due to the enormous intricacy of the object matter, and the obscure manner of representation by Kant and his followers, the established results of that new logical discipline have not yet penetrated into the circles of cyberneticists and designers of computing machines. There are two reasons for it, the first one rather personal: you can't get to the bottom of transcendental logic without first going very well versed in symbolic logic, Aristotelian logic, psychology, psychiatry, and last but not least ontology (general theory of objects). Only now have mathematicians begun to take in symbolic logic; and rarely do they advance beyond the technique of logical plumbing.

The second reason is to be found in the as yet rather undeveloped state of cybernetics. This new and amazing discipline has not yet arrived at the level of problems where the procedures of transcendental logic put in their appearance. Consequently, the provisional neglect of Kant's reasoning about "transcendental" mechanisms in the mind has had no ill-effects – so far at least – on the progress of engineering of mathematical and logical computers.

However, the story is quite different with regard to the universal theory of mechanical brains. The present discussion of the question: can a mechanism really think? (in other words: can they have material thoughts *accompanied* by consciousness ?) is, a clear symptom of a basic confusion. Here transcendental logic might be helpful. I shall therefore develop on the next few pages the theory of consciousness as established in transcendental logic. My exposition, incidentally, will avoid all original Kantian and Hegelian terms (with one notable exception) and will adapt itself to the technical requirements of modern cybernetics. (The scientific reader who might take exception to the simple terms and primitive similes employed in this article is referred to my book : "Elements of a New Theory Thinking in Hegel's Logic"; Leipzig 1933). This book presents the same theory minus the cybernetic viewpoint with the necessary scientific rigour.

To begin with: the skeptics who insist that mechanical brains are intrinsically incapable of conscious thought are wrong. This can be confidently asserted, because, transcendental logic is capable of a satisfactory definition of what human consciousness really is *and how it woks!* With this definition introduce our one and only original transcendental term: consciousness is *reflection in-itself.* But what does that mean ?

Everyone knows the simple phenomenon of reflection. You have only to look into a mirror in order to see the reflection of your face. You, also see an extended series of reflections, when you watch a movie on a theater screen – in which case,

it is the white screen that reflects (throws back at you!) the changing images of the film. Surely, then, the screen reflects events; but nobody who is in his right mind would say that the screen has consciousness. For the screen does not know what is happening. The light that bounces from it is reflected into *our* eyes, and only we, the audience are conscious of the events of the film. Therefore those reflected events are not reflection-in-themselves.

The story would be entirely different if the light were not thrown back at us, the audience, but were instead reflected back upon the *projector* and its optical process of projecting the images against the screen. Then we would not be aware of the whole affair. All possible consciousness would then be vested in the optical events going on between screen and projector. However, let us not stretch this simile too far. It only serves to give an approximate idea of what transcendental logic means when it, uses the term: reflection-in-itself.

Now: consider your own consciousness, a sensitive "screen". This "screen" receives, through your sensorial system messages from the outer world. Neuronic impulses coming from you our eyes, your ears, your skin, your muscles, etc. impress themselves upon that "screen" and are *reflected*. But this reflection is not thrown back at the world-system from which it came (as in the case of the mirror or the movie screen). Instead, it is thrown into a deeper recess of your brain, turns around and appears a second time on your brain-"screen", superimposing a second reflection on the first. This *second appearance* establishes the miraculous phenomenon which we call "consciousness."

Let's illustrate this process with a simple example: you are aware of a flower. This object of the outer world sends messages through your senses to your brain-"screen", where a picture of the object is formed. The picture bounces off the "screen" as unconscious message: "a rose." Then it goes to some other part of your brain, and returns to the first place with the superimposed content "acknowledged". Now the image on your brain-"screen" has a functional depth-dimension which is expressed in the statement: *I see* a rose. The original message "a rose" does not establish consciousness, because it is a simple reflection, not unlike the one in the mirror; but *the returning message does*, for it is a reflection- in-itself – or as we moderns should rather be inclined to say, it is a reflection upon itself.

Now, it is obvious that we should be able to design consciousness technically if we could find out what happens to the message after it has been first received on the screen of our brain and before the later moment, when it returns to it with the stamp "acknowledged" and produces consciousness by its second impact on the screen. (Incidentally, the time-lag between the two moments is so small that it is unobservable by the normal method of introspection.)

Fortunately we know what happens to the message during this reflexive interval and it is this theory of the brain processes during the round-trip of our message that is called "transcendental logic". Our verb "to transcend" means "to go beyond" (Roger's Thesaurus, 303) and we are entitled to ask: what did Kant mean by using this term? To go beyond ... what? The answer is simple, but quite unexpected. Till the publication of "The Critique of Pure Reason", philosophers and scientists had entertained the following ideas about the origin of consciousness: they said, our

mind is like a jug into which you pour water. The water while it is poured is in a rather chaotic state. The jug, however, stills it, and forces the fluid to adopt its own hollow form. According to this theory, then, our consciousness is a system of hollow forms into which are poured all the sensations, impressions and stimulations which our nerve system transmits from the outer world. But these transmissions arrive in a rather chaotic state. They become conscious only by being submitted to a forming and ordering mechanism which gives them their final (i.e., conscious) shape. The scheme is so simple, and moreover – as far as it goes – absolutely correct, that even nowadays 99.999% of all people adhere to that explanation.

They say: our mind has two fundamental components, namely contents and forms, and if the two come together the result is consciousness. If we talk about the universal reservoir of possible contents of our consciousness, we say: "material world"; if we talk about the jug these contents are poured into, we say: "formal logic".

The first description of our forms of consciousness, and how they work in order to shape the incoming material, was originally given by Aristotle. Since then, "formal logic" and "Aristotelian logic" have been historically equivalent terms. However, the "jug" Aristotle described was comparatively small. The Stoics, later, enlarged it a bit and since the introduction of Boolean algebra it has been discovered that all our previous conceptions about the size of our "jug" have been ridiculously conservative. The "jug" is still growing. Now it is usually called: mathematical logic; but it is still of course the same venerable vessel of ancient origin: a *formal* logic – meaning the theory of a mechanism that forms and orders contents.

The only trouble is: if you pour water in a jug, this vessel does not become water-conscious; and if you charge a battery the battery does not become electricity-conscious.

This did not disturb the philosophers of the Platonic and Aristotelian tradition. They said: it is different with man. Man has a soul. The inanimate object has not; and you need in addition to that synthesis of forms and contents a Self that watches that synthesis, thus finally producing that miraculous phenomenon: consciousness.

To the scientist, of course, the introduction of the term "soul" is nothing but a very polite way of saying: there is something in addition to this form-and-content business, but we don't know what it is. It was Kant who in his "Critique of Pure Reason" eliminated the concept of "soul" from the theory of logic (earning him an indictment of "atheism") and who stated that beyond the mechanism of formal logic there is in our brain a second mechanism which works on entirely different principles. It does not *form* messages any more but *carries* them through processing stages and finally returns them to the original "screen", the identity level of the formal logic. Insofar as this carrying capacity which transports the messages first *beyond* the screen is the most outstanding feature of this second brain-mechanism, Kant called the theory of it "transcendental" logic.

This theory is capable of demonstrating that if the message "a rose" is carried beyond the original "screen" and processed in a well defined manner, then the

concepts "I" and "perception" are added. These additions, however, do not by themselves produce consciousness. They are *pre-consci*ously attached. Only when the thus modified message returns to the screen is consciousness actually produced.

This happens in the following way. The returning message does not return to all parts, of the screen, but only to two sections of it, called "memory" and "identification" (the classical axiom of identity). The memory still retains the original pattern (unconscious) :

"a rose" ;

on which is superimposed (unconscious) :

"I see a rose".

Identification now produces a *confrontation* by attempting to establish a one-to-one correspondence relation between the original pattern and the enriched second message. This does not work! It turns out to be impossible to establish, by confrontation, a one-to-one correspondence between "a rose" and "I see a rose". The first part of the second sentence: "I see ..." overlaps. In other words: the reflection-in-itself produces, something that cannot be identified with the mere content "a rose". A tension of meaning is created – a tension between identity and non-identity. And this is the moment when consciousness and conscious thought comes into existence. No mysterious soul is necessary to explain the workings of consciousness. It should, however, be stressed that transcendental logic demonstrates only that consciousness is a mechanical process. Consciousness is that state in which a person is aware of the objective world. In other words: consciousness is equivalent to being aware of objects located outside the system of awareness.

It is quite a different story whether self-consciousness is also mechanical and therefore artificially reproducible. Self-consciousness is not awareness of objects, but of awareness itself. It is awareness of (awareness of objects). Transcendental logic, in its present form at least, does not extend over the range of this new problem. If I am permitted to voice an opinion, I should like to say that I do not believe that self-consciousness in its full dimension will ever be reproducible. Maybe carefully isolated fractions thereof – but that is the most we should hope for, and even this very limited ambition may find some realization only in a very distant future and on a considerably higher historical level than we are living on now.

However, the theory of transcendental logic enables us to discuss intelligently the question whether mechanical brains may have consciousness (not self-consciousness, mind you!). From what we have said on the preceding pages of this article two conclusions can be drawn. First: genuine mechanical brains which would deserve that name, would have consciousness. Because, if consciousness is a process whose workings can be described by an exact logic, then it can also be reproduced mechanically. (After all, symbolic formulae are nothing but a mechanism projected on paper.)

The second conclusion is: none of the present designs of logical or arithmetical computers, be it digit or analogue machine, can ever be brought to such a

perfection that it may eventually attain consciousness. Consciousness simply does not lie in the direction where progress is at present being made. Our resent designs try exclusively to solve the problem of how to reflect information upon a mechanical system. This problem has two technical aspects:

a) how to transpose physical events (e.g., electrical impulses) into patterns of information,

b) how to use these patterns as "motives" for operational procedures that follow the formal laws of identity, forbidden contradiction and excluded middle.

But this is simple physical reflection, modified only according to the peculiar properties of information. It is not, and will never be, reflection-in-itself. Technical aspect (a) repeats the sensorial transmission of messages by the human organism to the "screen" within the brain. Aspect (b) repeats the ordered arrangement of the data upon the screen. It is quite possible that these mechanisms within the present types of calculators shall finally be so perfected that they surpass beyond all imagination the functions of the human brain which they parallel. Nevertheless, none of these technical wonders shall ever have consciousness because that "carrier"-mechanism is lacking that permits the information to bounce off the screen and return to it in a modulated manner for the purpose of "confrontation".

It follows that in order to produce consciousness within a mechanical brain, entirely new designs will be necessary. These novel designs will contain as sub-systems the present type of calculator (although in a considerably improved variety) with a very significant additional feature: these sub-systems must perform their own coding. The reason is obvious: as long as man does the coding, the logical principles according to which the calculator is working are located in part outside the machine: are represented by the actions of the person who does the coding. As long as that is the case, the calculator is not in the possession of vital information (retained by the coder) that is needed to whip the information into proper shape for the transcendental "carrier" operation.

As this point is of utmost importance, let me rephrase it. In the present calculators only a small fraction of the whole system of formal logic is incorporated into the electronic mechanism. The greater part of it is still handled by the human operator of the machine – a factor in the operation which, of course, does not turn up on the "screen". This means: the logical information the "screen" receives is incomplete. And you cannot reflect an incomplete system upon itself! (One of the main theorems of transcendental logic.) Thus the designer is forced to build the whole system of formal logic into any calculator which (or is it from now on: who?) is supposed to think for itself (himself?).

For the time being, *we*, the humans, operate the calculators. But if a mechanical brain possesses consciousness, it is supposed to operate itself. Autonomy of action is one of the necessary prerequisites of any form of consciousness. A plant is rooted to the soil. It has no freedom of action, and we are fairly certain it has no consciousness either. An animal has freedom of action in the world, and there is no doubt that the animal organism is endowed with consciousness.

We humans are gifted with two forms of consciousness. On the lower level our powers of awareness are purely animal. However, the human organism has developed a second system of self-determination, the rational mechanism of logic. And as much as the animal organism of the body, is necessary for a *conscious* orientation in the world in terms of physical action, so is the rational mechanism of logic. the necessary vehicle for *conscious* orientation and *action* in the realm of abstract thought.

However, you always need the whole system for autonomous, (conscious) operation. There is a story in European folklore of a blockhead, who was a skin-flint to boot, who argued: "I need my horse for running. For running he needs only his legs. His head doesn't do me any good. He needs it only for feeding. If I chop off the head I shall save the food, and still have the legs for running." It seemed quite a convincing sort of argument. So this blockhead cut off the head of his horse and got the surprise of his life when the four horse legs refused to run afterwards.

In abstract terms: conscious action demands a complete system; and a horse without a head is not a complete organic system.

Similarly it is quite impossible that a mechanical brain can ever consciously think without having a complete system of logic built into it. Our present designs contain only fragments of logic. None has the concept of identity as systematic integration of all (potentially conscious) procedures built into it. The symbolic formula for the definition of "identity" is:

$$(x = y) =_{Def} (f)\ [f(x) \cong f(y)]$$

This expression can be read: Two objects (of our consciousness) "x" and "y" designate semantically the same object in the world) if any two corresponding statements "f(x)" and "f(y)", which contain these symbols in corresponding places, have equal truth-values.

This formula shows: it is not sufficient to build the calculus of proposition and the calculus of classes into a mechanical brain which is supposed to think for itself. You have also to design the calculus of functions into it, and not only the calculus of simple functions, but also the extended calculus of functions (of functions). This latter task is beset with enormous difficulties. The functional calculus of second order implies the logical as well as the semantical antinomies. It can be shown that a mechanical brain which really possesses consciousness must be subject to the pitfalls of (human) thinking in exactly the same manner as we are.

At present we have neither the theoretical nor technical implements to complete such a task. A logician who knows his higher calculus functions might well say that the difficulties are absolutely unsurmountable. I should agree with him – if we were called upon to design an exact replica of the human system of thinking. But we are not. The human system of logic covers not only consciousness, but also self-consciousness. The designer of the mechanical brain, on the other hand, has only, to deal with the problem of consciousness. That permits certain simplifications which the logician who describes the system of thinking cannot afford. Therefore, I think, it should finally be possible to replace the human

system of thinking which operates the calculator with a built-in-system of robot-thinking that operates the brain internally.

This mechano-logical operator would be included in the "transcendental" carrier-system. A mechanical brain endowed with consciousness would consequently not be a simple, epistemologically homogeneous mechanical system (as the present calculators are – and you'd better revise your ideas of what is "simple"!) but a very complicated system of systems with entirely heterogeneous modes of activity. A primitive drawing might help, as shown in the accompanying diagram of the transcendental system (see diagram_1).

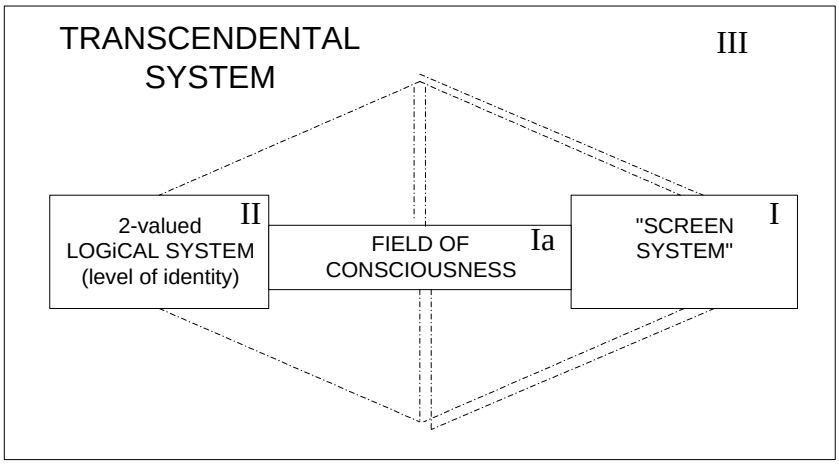

Diagram_1: Transcendental System

None of the present calculators has a design beyond system (I). In fact, our modern machines fill only a very small fraction of (I), with some tiny scraps of (II) thrown in for good measure. Both systems, (I) and (II) share equally in the field of consciousness (Ia). But before consciousness can be mechanically created, all messages arriving at (I) are first carried into the "transcendental" system (III). From there they return to (I), resp. (Ia) This roundtrip is effected by a feedback mechanism. This is a discovery of Hegel, who describes, consciousness as a logical feed-back in his "Phenomenology of the Mind" (pp. 193-221, and 549-564 of the edition of 1928). Our dotted lines indicate this primary feedback mechanism. But there is a secondary feedback which connects (II) with (III) and (I). The feedback connection between (II) and (I) should be indirect, especially as there is a direct connection through (Ia).

This drawing should, of course, not be interpreted as a blue-print, however remote of the ultimate technical reality. It is merely an illustration to show how little has been done as yet towards the realization of the idea of a genuine mechanical brain.

You can see that the systems (II) and (III) may be thought to represent what, in the mythological language of mankind, is called a "soul". If you just say soul you mean system (III); if you speak about a "rational soul", then you add (II) and (III) together.

The general concept of system (III) dates as far back as Plato's dialogue "Theætetus". In order to demonstrate that consciousness demands an integrating unit, Plato uses the example of the Trojan horse. Inside this horse were seated

many Greek heroes, like Ulysses, Diomedes; and others. But although there were brain functions going on "inside" the horse, this wooden monster did not derive any consciousness from them. Accordingly, young Theætetus is told: "It would be a singular thing, my lad, if each of us was, as it were, a wooden horse, and within us were seated many separate senses, since manifestly these senses unite into one nature, call it soul *or what you will*; and it is with this central form through the organs of sense that we perceive sensible objects."

There is little doubt that our present "thinking" machines are hardly more than wooden horses.

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

vordenker
ISSN 1619-9324

How to cite:
Gotthard Günther: Can Mechanical Brains Have Consciousness?, in: www.vordenker.de (Winter-Edition 2005), J. Paul (Ed.), URL: < http://www.vordenker.de/gunther_web/mechan-brains_conscious.pdf > — originally published in: Startling stories, Vol. 29, No. 1, New York, 1953, p. 110-116.