Δαίμων 2 - 1990. Págs. 125-138

The Order of Thought
Wittgenstein on Artificial Intelligence and Brain-Processes

ALBERTO EMILIANI (Università degli Studi di Bologna)[*]

ABSTRACT

This paper is *not* devoted to analyzing Wittgenstein's claims about machine-thought but to clarifying and expanding an argument of Wittgenstein's about the non-mechanicity of thought. According to my reading of *Zettel* 608, such a feature of thought should not be merely accounted to the contrary, it should be argued for by showing that an analysis of our brains does not provide any account of what thought consists in (a conceptual case).

The point is quite radical: a neural structure does not define a concept, but a concept defines what neural (or otherwise physical) structures would answer to it. Concepts are therefore not mere "emergent properties" of neural frameworks – unless by "emergent property" we mean a property which does not *proceed* from its underlying *substratum*.

On these grounds, there is no need for thought to be mechanical at all, for the order of thought has not to answer to a physical (or even to micro-logical) order which is mechanical in nature. The inspection of a case of vagueness is intended to illustrate the point at issue.

I

An investigation about Wittgenstein's ideas on Artificial Intelligence should first answer the question whether Wittgenstein ever had any ideas on this subject.

[*] Facoltà di Lettere e Filosofia. Dipartimento di Filosofia. Via Zamboni, 38. I-40126. Bologna.

Nothing which Wittgenstein says is explicitly concerned with what we call Artificial Intelligence. He never mentions problems concerning the notion of a program – nor he distinguishes between such theses as "strong Artificial Intelligence" and "weak Artificial Intelligence" in the sense of Searle 1980. Putnam's "functionalism" (cf. e.g. Putnam 1960 and Putnam 1967) and Searles theses about the ultimately physical nature of intentionality (cf. Searle 1980 and Searle 1984), which play an outstanding role in the present debate, were obviously unknown to him. Therefore the question arises whether what Wittgenstein has to say about minds and machines is at all relevant, but from a historical point of view.

The answer is that Wittgenstein has quite a lot to say about these matters. On the one hand, he has been viewed as an indirect (and partial) supporter of the Artificial Intelligence theses. Such an assessment rests, in my opinion, on a wrong intepretation of Wittgenstein's philosophy. However, the mere fact that interpretations of this kind have been put forward (cf. e.g. Obermeier 1983 and Wilks 1976) shows that there is something to be discussed. It should also be noticed that Dreyfus (one of the opponents of AI) explicitly acknowledges Wittgenstein as a source of inspiration of his ideas (Dreyfus 1979).

On the other hand, the problem whether an explanation of our concepts might be based upon the features (be they physical or formal features) of our brain(s) *is* present in Wittgenstein's investigations, and much which Wittgenstein says is relevant to it. In this sense, a thesis about (and against) Artificial Intelligence can be real off from his writings.

According to Obermeier 1983, Wittgenstein's philosophy is consistent with the main theses of Artificial Intelligence. The argument of Obermeier can be outlined as follows:

(1) In his "private language argument" Wittgenstein holds that there are no "privat objects" — or at least that the existence of private objects of whatever kind has nothing to do with "thinking". Ascribing thought to somebody, i.e. asserting that some body thinks, has only to do with his actions.

(2) This constitutes an indirect criticism of the mentalist arguments against Artificial Intelligence; the mentalist claims that machines cannot think because they lack the relevant internal states — a kind of state which is neither physical nor intersubjectively accessible. Therefore the relevant states are "private objects".

(3) If an ascription of thought is based upon what one may call "correct behaviour" then machines are also capable of this much; therefore the Wittgensteinian arguments against the possibility of a private language can be interpreted as arguments to the effect that machines can think. That machines cannot think is nothing but a common sense assumption which scientific research and wider cultural changes may overthrow.

The *prima facie* plausibility of such a line of thought relies upon a substantial misunderstanding of Wittgenstein's philosophy – especially of his reference to behaviour. That Wittgenstein was neither a "standard" behaviourist nor a behaviourist *"sui* *generis"* turns out clearly from an analysis of his last writings (especially of the writings about the philosophy of psychology, partly embodied in the so – called Part II of the *Philosophical Investigations*). Prof. G. Luckhardt (cf. Luckhardt 1983) gave an insightful account of the matter which I refer to for further discussion.

For the time being, there is no need to engage in investigating behaviourism, since the hasty intepretation given in (1)-(3) is contradicted by some ten remarks included in Zettel (Zettel, 603-615 and following). There Wittgenstein very explicitly asserts that the physical processes taking place in our brain are completely unessential to an explanation of what thinking and other "psychological phenomena" (Zettel 609) consist in. Now, if the crucial thesis of Artificial Intelligence is that what constitutes our thought is the computational structure of the processes taking place in our brain(s), then this is sharply denied by the Zettel remarks mentioned above. However, if that thesis is understood in the somewhat weakened form that machines can think, irrespective of what thinking is supposed to consist in, then Wittgenstein's position is not equally explicit. All the same, it can be argued that Wittgenstein rejects the theses of Artificial Intelligence, both in their firsi and in their second formulation.

In the present paper I shall only concern myself with the first thesis of Artificial Intelligence that machines can think *because* thought can be reduced to the computational structure of our brain-processes and such a structure can (in principle) be implemented in a computer. It is a thesis about thought and it is part and parcel of the mind-body problem. It is this thesis that constitutes the pivot of the debate about Artificial Intelligence. As far as I know, the second thesis has never been stated before; it is in fact quite uninteresting unless the first has been proved to be fallacious. A neat criticism of the second, thesis can however be provided (a very sketchy account of it is given at the end of section IV).

II

The question which Wittgenstein tries to answer is the following: can the nature of thought be revealed by an analysis of our nervous system? Admittedly, Wittgenstein does not distinguish a conception of our brain seen as the implementation of a program from a brain seen as a system of physical processes. Is such a distinction all that crucial? I believe not. For Wittgenstein's point is that nothing which happens in our brain can be of any interest to a conceptual investigation. Now, whether that which happens in our brain has to be understood from a "formal" (strong Artificial Intelligence) or from a physical (Searle) point of view is beyond this point, since Wittgenstein simply holds that *nothing* happens in our brain which is essential (or even relevant) to an analysis of concepts.

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be

> impossibel to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a *system* of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? The case would be like the following — certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced — but *nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that comes out of it — this can only be done from the *history* of the seed. So an organism might come into being even out of something quite amorphous, as it were causelessly; and there is no reason why this should not really hold for our thoughts, and hence for our talking and writing.
>
> (*Zettel* 608)
>
> It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them.
>
> (*Zettel* 609)

At first sight, Wittgenstein's statements look exceedingly irrationalist — perhaps even unreasonable. My opinion is that Wittgenstein actually goes too far, and mixes up two different kinds of arguments, one of which is both powerful and reasonable while the other is relatively poor.

Let us start from the (relatively) poor one. It is based on what one might call the constitutive weakness of science. It runs as follows. There is no *a priori* argument to the effect that cases as that of the seeds, or that of an organism coming into being "as it were causelessly" from something quite amorphous, are impossible. Science cannot guarantee that impossibility. Actually, science cannot exclude that new kinds of phenomena will ever be discovered which cannot be adequately treated through causal investigation. It is logically possible that certain kinds of phenomena are subject to no laws, no regularity at all. This is as old as Hume. Modern philosophy insisted that the aim of the scientific enterprise does not consist in revealing hidden features of reality. Things such as regularity, or conformity to causal laws, are instead presuppositions of the scientific investigation; they are structural features of our knowledge.

Now, the hypothesis that seeds of the Wittgensteinian kind exist (or the hypothesis that I might have no brain at all) is not logically impossible, not any more than the hypothesis that the Sun will not rise tomorrow. However, this does not show that such a hypothesis is natural, nor is its philosophical purport made plausible by the mere circumstance that such a hypothesis is not inconsistent. For the only conclusion which such an argument supports is in fact that it is not *necessary* that our thought corresponds to a system of neural processes.

It should be noticed that the mere possibility of a failure of causal explanation *might* have been a good starting point for an argument. Since thought is not necessarily connected to a system of neural phenomena causally related to each other, a system of that kind cannot be seen as a constitutive feature of thought — nor, a *fortiori*, as "what our thought consists in".

The argument might have been further pursued by showing that, since every scientific explanation presupposes a conceptual framework, no scientific explanation might sensibly attempt at explaining the conceptual framework itself. Such an attempt would lead to paradoxical consequences, as many (e.g. Husserl, *Prolegomena*) have shown.

Be this as it may, Wittgenstein does not endorse such a line of thought. No supposition seems to him *more natural* than a failure of causal explanation. His example about seeds is very clear. He says that "*nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that comes out of it"— nor from those of the seed from which the plant comes out. Which amounts to assuming that it is impossible to trace the plant back to the seed *via* a sequence of changes causally determined. Analogous remarks apply to the example about remembering (*Zettel* 610). "Why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored up there *in any form*?"

The need for a non-causal explanation is closely connected to a failure of causal explanation. Therefore the argument is poor. The mere logical possibility that the Sun will not rise tomorrow is not a good reason for doubting that it will actually rise. My intention is not to state that Wittgenstein's *conclusions* are untenable; only that his argument, based (as it is) on the possibility of causal failure, is extremely weak.

As a matter of fact, the second argument is relatively independent from the first, and from any hypotheses about failures of causal explanation. The point of the argument might be summed up as follows: even if all that happens (including our utterances) were causally connected to what happened before, a causal explanation of our brain-processes and behaviour cannot possibly account for the logical structure of thought. There is an order of thought which is independent from any physical order. If such a statement holds good, then another, more subtle, kind of failure of causal explanation is brought to light; not a failure in its own domain —a lack of causal regularity— but a failure consisting in the impossibility of explaining something by means of causal relations holding between micro-phenomena.

Now, if the order of thought is actually independent from causal order, *then* it might be natural to think that any causal link between neural events is of no importance, after all: for a failure in that domain would not affect the order of thought.

From this point of view, the "supposition" of Wittgenstein is, if not natural, not unbelievable either; this is perhaps the reason why the two arguments conflated in his mind.

Another possibility is that Wittgenstein put these remarks apart, instead of including them in a major work, simply because he felt that they were not good. The nature and purpose of the cuttings found inside a box and published as *Zettel* is unknown; thinking of the box as a waste-basket is certainly unfair, but I find it very likely that many of the cuttings hould have been extensively revised.

My primary purpose is neither historical nor exegetical. I will not further discuss whether and to what extent some ten remarks from *Zettel* are to be considered finished and refined. My aim is to find out the rationale of what I called "the second argument"; whether it is acceptable and why. To this end, I shall discuss some problems connected to the independence of the order of thought, providing arguments of my own in defence of it and reducing to a minimum the references to other features of Wittgenstein's phylosophy.

III

The Wittgensteinian argument of Zettel 608 allows two different readings. The idea that the order of thought is independent from the order of brain-phenomena might be traced back to an analysis of apsychological concept of thought. Alternatively, the independence of the order of thought might be argued for by appealing to the *logical* features of thought, i.e. by trying to show that the logical structure of thought cannot be identified with a structure of brain-processes. In the former case, the arguments refer to a psychological notion of thought, whereas in the latter a logical notion of thought is concerned. I will try to show that there is a link between the two readings. However, the pivot of my discussion is a logical, rather than a psychological, notion of thought.

Both the supporters of Artificial Intelligence ("strong Artificial Intelligence" in the sense of Searle) and Searle share a crucial assumption: that our thought *consists in* (alternative formulations are: that the essence of thought is; that thought can be reduced to) the neural processes which correspond to thought in our brain. That "we think with our brain" might be their *motto*.

On the one hand, as far as Searle is concerned, my assertion needs no proof: that our thought consists in the system of physical processes taking place in the brain is, in a nutshell, the whole point of Searle.

On the other hand, as regards Artificial Intelligence, the idea is that thinking — feeling, doubting, remembering, and so on — can be analyzed in terms of the formal structure of the processes which take place in our brain when thinking, feeling, etc. The physical substratum of such a formal (computational) structure is not essential; what is essential for thought being present is that there is a substratum whatsoever which implements the correct structure. In a sense, it might be stated that thought is not only what happens in our brain: it is all which happens in any substratum which shares the structure of our brain; or, in a wider view, that it is all which happens in such substrata and in all those substrata which share a formally equivalent structure. In any case, our thought consists in what processes (understood from a computational, instead of physical, point of view) which happen in our brain while we are thinking.

This point of view is sharply opposed to by Wittgenstein "One of the most dangerous ideas for a philosopher is, oddly enough, that we think with our heads or in our heads". (*Zettel* 605; cf. also PI 361).

If there is here a danger, then it is one which many are willing to run. But Wittgenstein's warning should be taken seriously. Grave misunderstandings and a good deal of confusion lie behind the claim that thought is essentially what goes on in the thinkers heads. Strange as it may seem, quite a lot of what Wittgenstein says against private languages can be turned (with minor changes) against an idea of thought as a system of head-events (cf. note 2). Actually, the mentalist, the "neural materialist" (as I would like to label Searle) and the supporter of Artificial Intelligence share much more than they might be supposed to. They agree that thinking consists in a hidden process. A sort of the beetle in the box of PI 293.

Thought is commonly ascribed to people without any previous analysis of their barin-conditions. Such an analysis is presently impossible — so complicated that it cannot be successfully attemped. But let us suppose that it is possible, and that somebody, say NN, who is considered as a competent speaker — thinker — does not satisfy the brain requirements for being a genuine thinker. Wat kind of conclusion should we draw? That we were wrong in saying that he really was a thinker?

What is funny is that none of us knows whether his own brain conditions are of the right kind. Might it turn out that I, or the readers of the present paper, are not thinkers? And what is the sense of such a hypothesis?

Assume that a closer analysis shows that many people do not have brain-conditions of the appropriate kind. Should we still say that they are not thinkers, or rather that our theory of thought was wrong?

Let us go back to NN, who thinks but has not the right sort of brain. The quarrel between mentalists and supporters of Artificial Intelligence presents us with many thought experiments having to do with machines which do have the correct structure but, owing to some reason or other, are held not to be able to think. I want to put forward quite a different example — a different "thought experiment". If a strange entity — ectoplasmic in nature — came to the Earth and spoke with us, maybe making funny jokes, or even writing best-sellers and winning a Nobel Prize for literature, what should we say? It certainly has no brain: does it think? Artificial Intelligence supporters might assert that the entity must have the correct formal processes *somewhere*; well, where? Mentalists would say that the entity must have the correct inner experiences; but how could we know that its inner experiences are comparable to ours?— as a matter of fact, the creature is very likely to have quite different inner experiences. Asserting that such an entity, since it lacks the correct "things", does not think, would be more than counter-intuitive: it would be blindly dogmatic.

Now, my aim is not to argue that such creatures exist. But, if the hypothesis of a creature without any brain, structured processes and inner experiences (of the right kind) is not the hypothesis of a non-thinker, this shows that the concept of thought has little to do with brains, processes or experiences; it has to do with what the cerature *does*: with the way in which it *acts*.

Problems of this kind arise if the status of the investigation about thought is unclear. The analysis of brain-processes is an empirical matter, whereas the claim that thought is nothing but this or that is essentially logical. The theses of Artificial Intelligence are based on an assumption which is logical in character; that is, that the regularity of our ascriptions of thought *must* correspond to the presence, somewhere in our head(s), of the contents of thought, *and* that the latter must in turn correspond to, and ultimately consist in, a physical regularity — or a regularity in the structure of our brain-processes [1]. The plausibility of such a thesis partly depends upon the fact that similar kinds of assumptions proved to be successful in many cases; e.g. in the case of sugar. Our regularity in calling this and that sugar corresponds to a physical regularity in what sugar is, a regularity in the molecular structure of sugar. Is the case of thought similar to the case of sugar?

Arguments aiming at showing that ordinary ascriptions of thought are based upon recognition of certain actions are meant to discard the first part of the assumption— the correspondence between thought and the "inner presence of the contest of thought" [2]. However, other and perhaps more convincing arguments can be provided against the "logical part" of the assumption, i.e. that the content of thought can be reduced to a structure of brain-processes.

According to the second part of the assumption, an analysis of brains provides an account of what thought consists in: it can explain what thought consists in. But it is clear that may different brains — that many differente brain-programs — might accomplish identical results. Let us consider a very simple case of thought-event; multiplying numbers up to 100. There are quite many (really infinite) programs which could accomplish this; and there are as many physical and brains (partly differing in structure) which could accomplish the same result. I do not mean that different implementation of the same program in different computers are actually different machine code programs. I mean that infinite programs are possible even if a "program" is intended as a source code expressed in a high level language. (From now up to the end of present section I will drop my double mention of physical brains viewed as substrata of a program: I will only refer to programs. However, what will be said of programs also applies, with minor changes, to brains physically considered).

[1] The present-day discussion about Artificial Intelligence is gradually shifting towards a discussion concerning the status of logic. It was a debate regarding thought – whether a machine can or cannot think. It became a debate concerning intentionality and understanding (cf. Searle 1980 and the wide debate which this essay arouse) and eventually a debate about *language* (whether the semantical features of language, i.e. meaning, the *content* of thought, can be reduced to a computational structure. Cf. e.g. Searle 1984).

[2] Arguments of this kind are frequently found in Wittgenstein's works (especially where he is concerned with "private languages"); an extensive account of these can be found, e.g., in Norman Malcolm's essays concening this subject (cf. Malcolm 1986). Tr. brief outline, the arguments concentrate upon the following point: we only ascribe thought to human beings or to what behaves like a human being. To think is therefore not a matter of brain-processes but a matter of behaviour and expression.

Program A takes the first digit of the first number and multiplies it by the first digit of the second number — it looks for the result in a table, stored somewhere. Then it goes on as usual, summing the partial results, etc.

B translates the numbers into their binary form, then effects a binary product and retranslates the result into decimal notation.

Before multiplying any two numbers, C always checks that 2+2=4. Under any other respects, it is identical to A (there are of course infinite kinds of C's).

D looks up the result in a table, where the results of all the products of two numbers less than 100 are stored.

That many versions of each one of the A-D's can be given even within the same computer language is a matter of fact: but I will not dwell on this point. Instead, let us ask whether one of these sequences of processes could be considered as "what multiplying two numbers essentially consists in".

Would it not be possible that the structure of what happens in the brains of different speakers while they multiply corresponds to different patterns of multiplication? Would it not be possible that what happens in the brain of, say, A' corresponds to the program A, etc.? Of course, much depends on the way in which I have learnt to multiply numbers. Much also depends on my individual attitudes, states and dispositions. Nothing is more reasonable than to suppose that I, for example, have the mental image of some kind of a root (say, a carrot) whenever I think of a square root (the example is Husserl's). Or that I have the mental image of a "X" whenever I perform a multiplication; or that this only happens in some cases, and that I have different images in different cases, depending on my mood. Now, there will be some additional brain-process going on while I think of the carrot etc.; and analysis — a sort of trace — of the brain processes which are going on while I multiply will also include those brain-processes. I think it very likely that similar differences between individual multipliers exist; in any case, the possibility of such a difference is obvious.

Thus, there are many (actually infinite) ways for performing the multiplication correctly; that is to say, there are infinte structures which accomplish the same result. Now, on what basis should we pinck up a single structure and say that it is the correct structure, i.e. what multiplying two numbers "really consists in"?

It might be answered that since different structures — patterns, schemes — occur in different cases, we actually have *different thoughts* and not the same thought. But in this case, should not the same be said for every other concept, or mode of reasoning? For many competent users of the concept — competent thinkers — might in fact "implement" different brain-programs. And, if people always think different thoughts, how can they understand each other? Or is our understanding, after all, fictitious?

Therefore, the supporters of Artificial Intelligence are forced to choose: one program (scheme, pattern) or another must be taken as what a certain thought consists in. But any choice would have extremely unreasonable consequences. If C' (that is, a man whose brain-processes for multiplication correspond to a program of the kind of C) always remembers some strange thing, e.g. a plant in his garden, before doing the calculation, and if he is unable to do the calcualtion otherwise, should we then say that he is not really calculating? And what if his brain takes a longer way — our tracing shows many inconclusive roundabouts?

The crucial question is a different one. In relation to what are C's roundabouts inconclusive? In the same sense in which another scheme, or pattern, of multiplication is "simple" or "elegant". Now, on the one hand, several alternative patterns can be simple or elegant. On the other hand, we *already have* an idea of what we would be willing to accept as a simple or elegant pattern of multiplication. It is on the basis of such a pre-existing idea that we judge upon the "simplicity" or elegance of the proposed pattern of thought. The idea of what it is to be a good pattern of "multiplicational thought" must be already present. Had we no idea of this sort, then any possible pattern would be right — provided that it gave the correct results.

Only if such an idea is present we can say that the alternative brain-programs attributed to different thinkers are, after all, equivalent: that this is closer to the ideal scheme — more likely, from one or another of a family of equally good schemes. It follows that the structure of our neural processes does not *define* what multiplying correctly is — it does not define what thinking of a certain concept is, etc.; such a structure must instead be seen from the standpoint of the conceptual structure of multiplication — of the concept etc. which is already given. When we analyze a program, be it a brain-program or a computer program, we cannot but see whether it *answers* or it does not answer to a conceptual structure of this kind.

What the argument shows is that the mere existence of a program, the fact that the output of the program is the correct verbal behaviour, is not a proof that that program grasps, or embodies, the correct logical structure of the fragment of natural language which it is intended to cope with. On the contrary, such a structure cannot be identified with a program; it can be identified with the structure of a programa *if* that program embodies the relevant structure. Which is by no means necessary, if we think of programs such as C or D. A higher parameter of comparison is needed.

This simple fact is of the utmost importance for any attempt at evaluating the theses of Artificial Intelligence. For one thing, it shows that the essential features of thoughts can only be found out by means of conceptual investigation. In addition to this, it shows that there must be a level, an order of thought, which is independent from the order of the physical phenomena happening in our brain; and *also* from the order of the formal, or computational, processes, to which those phenomena correspond [3].

[3] A supporte of Artifical Intelligence might now reply that the argument does not concern thought, understood as a psychological phenomenon, but the content of thought, that is, the logical aspect of thought. The order of thought which I am speaking about – the objection runs – is a logical order rather than a psychological one. I shall not try to account for the nature of the relations holding between logic and psychology. I shall limit myself to two brief arguments.

In the first place, my having a thought (the "psychological side of thought") should be someway related to the content of my thought (the "logical side"). If the content of my thought cannot be reduced to a chain (or to a structure) of brain-processes and, this notwithstanding, an attempt to reduce my thougth(s)

Multiplication is but an example, and one which very easily allows a translation in terms of a computer language. The conclusions which may be drawn from the argument stated above are deeper than the example might suggest. For, if the existence of a conceptual order which is independent from the order of brain (and computer) processes is shown, nothing forces us to conclude that every logical feature of language (and, *a fortiori*, of thought) can be grasped by a program — nor by a physical system.

That a program should (ideally) exist which grasped the essence of thought is closely linked to the thesis that the essence of thoughts consists in our brain-processes; this thesis implies that all the essential features of thoughts are embodied in the underlying system of processes. But the thesis is wrong: the program is not the esential structure of thought, but it may, or may not, embody such a structure. Therefore, the question may be raised whether "the essential structure" of thought may, in all the cases of thought, be embodied (and expressed) by a program of some kind.

In the case of multiplication we have many programs which express the main features of multiplying. Programs as A or B *do embody* the logical structure of multiplication. That analogous programs might be provided for *all* the "ways of reasoning", or *all* the concepts of our language is not a logical necessity, connected to the nature of thought (it would be so if the essence of thought consisted in the underlying structure of brain-processes); it is a hope of Artifical Intelligence (both of "strong" and "weak" Artificial Intelligence) — and a hope which, as a matter of fact, will not be fulfiled.

There are here many *degrees* of fulfilment. A limiting case is that of concepts (and generally of logical structures of fragments of natural language) which can be naturally expressed in a high level computer language (be it procedural or declarative). Multiplication belongs to this family of cases. Opposite cases are those of other framents of natural language (I think, e.g., of the "languages" of Religion, Ethics and Art) whose "computer translation" is utterly unaleg to express their features - it can at most simulate a correct liguistic behaviour, but the closest analysis of the program would be of no help in order to understand the concepts and the logical structure of those fragments. There are also many intermediate cases. I will discuss in some detail some problems connected with a case of "vagueness". In such a context, some of the notions referred to above (as the notion of simulation, and of a program embodying a structure) will be clarified.

to such a physical chain or structure is carried on, one gets the awkward result that the content of my thought is not present in my thought. In the second place, even a psychological concept of thought cannot survive dramatic gaps between the thinkers. I mean that, on the one hand, we may assert that, if C' has to think of a carrot whenever he performs a square root, then what he thinks is different from what A' thinks; but, on the other hand, thinking of a carrot must be univocal enough to allow an ascription of the *same* thought to other people as well; whereas in different people the carrot-thoughts might correspond (and very likely they do correspond) to differente brain-processes. Therefore, what makes all those thoughts into the same thought (the unity of the concept) has to be looked for in a logical space which is not the dimension of brain-processes.

IV

Let us consider Wittgenstein's example of the order "Stand roughly here". Such an order is *structurally* "inexact". The vagueness which affects this order is a logical feature of it. A closer analysis of this case might shed some light on the distinction between a logical order and a causal order.

"Stand roughly here". It should be noticed that the meaning of these words is not "stand in a sharply exact place which I do not explicitly express". What is meant is really, and only, "roughly here". My purpose is not to state that a machine will never be able to carry out, or even to give, an order of this kind. On the contrary, it is quite reasonable to think that programming a robot so as to enable it to perform similar orders should not be an impossible task (the heart of the problem is not what a machine can do, but whether our thought is mechanical). The real question is whether the program which makes the machine to behave in a satisfactory way embodies the logical features of the order, such in a way that an analysis of the program would allow us to grasp these features.

In order to be able to perform the order, the machine has to "know" that it has to stand still in a certain place, at a distance *m* from us, or at a distance *m* + or - some *n* *m* and *n* are obviously exact numbers. They might also be determined in a very complex way, or even by using pseudo-causal functions; but, having we to do with a program, those numbers have to be exact. Well, due to a "weighted" determination of *m* and *n* the machine behaves satisfactorily — or so I want to assume. Does this support the conclusion that the "true" logical structure of "stand roughly here" is the same as that of "stand in whatever position not further from me than *m+n* and not closer than *m-n*, for *m* and *n* determined so and so"?

Our vagueness is not a sophisticated, and perhaps "economical", form of exactness. "Stand rougdhly here" is not an ambiguous expression of an exact order; to understand the order does not mean to understand that one has to stand in whatever position within a sharply determined area. What I mean is that vagueness constitutes a proper part of the structure of the order; and this shows that a "mechanical" translation, far from revealing what the logic of the order is, has the effect to change (more or less significantly) that structure.

There are two different problems: one is technical and the other is logical. To the first, whether the machine can behave as those who understand vague expressions, my answer is affirmative. But the more significant problem is the second: whether by "teaching" the machine so to behave we also accomplish a reduction of the logical structure of vagueness to a "mechanical" structure, i.e. to a structure which can be embodied in a computer program. And in this case the answer is necessarily negative. In other words, a reduction to a program logic is not a means for clarifying or analyzing the logical essence of the concept; it is but a method for accomplishing a machine-behaviour which approaches as closely as possible to a pattern fiven in advance.

We will observe the behaviour of the speakers; we may perhaps carry out "experiments", soas to establish that standing still in a certain place complies with the order if that place is within a certain distance from the place pointed at when saying "here" (which, in any case, is another vague expression: they are countless, and the concept of "exactness" itself is vague). We might also discover that the distance which is in question is a function of certain variables, or even of a context. Once this has been done, we are perhaps able to program a robot which is able to carry out "inexact" orders, or to establish whether somebody else has correctly performed them: but we will have accomplished no reduction of vagueness to exactness. There is no reduction of vagueness to exactness; but it is possible to destroy vagueness.

I would like to draw two conclusions from what has been said. First of all, the possibility of building a machine which is able to comply with vague orders does not constitute a "proof" that the logical structure of that order "consists in" or "can be reduced to") the logical structure of the program. As a matter of fact, the program should embody several *ad hoc* adjustments which will enable it to cope with the standards of correct use. Such standards are given by patterns of human actions. In a way, what the program does is nothing but *simulating* a full understanding of the order, since the logic of the order is not expressed by the program itself.

The objection may be raised that a program may (in principle) be written which is the copy of our "brain-program". Therefore, if we are able to think then the program should also be able to do so. But the objection assumes that our brain is the *locus* of understanding; and this is the assumption which has been questioned up to now. For the point of the arguments presented above might in fact be stated as follows: that conceiving us as machines does not provide a full explanation of our thought — it only allows to see us as creatures which simulate thought.

In the second place, I want to argue that our actions set a model, an ideal which the program should be able to reach. If the model is such that it cannot be grasped by the program (i.e. such that the program cannot capture its logic) but only complied with by it, then the fact that the model is set by men is one of the constitutive features of the model, and not a merely contingent one. I am aware that a point like this needs further explanation and development; which I cannot do in this context. However, it should be emphasized that this line of thought is the pivot of the answer to what in the first section of the present paper I have called "the second thesis of Artificial Intelligence".

The existence of a concept, of a logical structure, sets standard for correct use. We may be able to satisfy the requirements imposed in this way; and usually we are able to do so. What brain processes take place in our head, what mental (more or less private) events take place in our mind, can be *means* by which the task is fulfilled; they are not the task itself; in many relevant cases the closest analysis of them is utterly mute on the nature of the task (as if we tried to find out what "being a King of Spain" consists in by examinig the legs, arms and head of King Juan Carlos).

Thinking, speaking, resembles jumping over a hurdle. There are several techniques of jumping; some people need more time; some people might climb on a chair in order to make their jump easier. In such different ways anyone strives after his aim. But if one tried to find out what the aim of jumping over the hurdle "essentially" consists in by analyzing the bodies of the jumpers, or the structures of their movements in the space, he would condemn himself to an endless and unfruitful research.

REFERENCES

DREYFUS, H.: 1979, *What Computers Can't Do, The limits of Artificial Intelligence*, Harper and Row, New York.

FREGE, G.: 'The Thought: A Logical Inquiry'. An English translation in P. F. Strawson (ed.), *Philosophical Logic*, Oxford University Press, 1967.

HUME, D.: *A Treatise of Human Nature*.

HUSSERL, E.: *Logische Untersuchungen*.

LUCKHARDT. C. G.: 1983, 'Wittgenstein and Behaviourism', *Synthese*, 56.

MALCOLM, N.: 1982, 'Wittgenstein: The Relation of Language to Instinctive Behaviour', *Philosophical Investigations*.

MALCOLM, N.: 1986, 'Mind and Brain', in *Nothing is Hidden*, Basil Blackwell, Oxford.

MAURY, A.: 1981, 'Sources of the Remarks in Wittgenstein's Zettel', *Philosophical Investigations*.

OBERMEIER, K. K.: 1983, 'Wittgenstein on Language and Artificial Intelligence: The Chinese-Room Thought Experiment Revisited', *Synthese*, 56.

PUTNAM, H.: 1960, 'Minds and Machines', in S. Hook (ed.) *Dimensions of Mind*, New York University Press.

PUTNAM, H.: 1967, 'The Mental Life of Some Machines', in H. N. Castañeda (ed.), *Intentionality, Minds and Perception*, Wayne State University Press.

SEARLE, J. R.: 1980, 'Minds, Brains and Programs', in *The Behavioral and Brain Science*, Cambridge University Press.

SEARLE, J. R.: 1983, *Intentionality*, Cambridge University Press.

SEARLE, J. R.: 1984, *Minds, Brains and Science*, British Broadcasting Corporation, London.

WILKS, Y.: 1976, 'Philosophy of Language', in E. Charniak and Y. Wilks (eds.), *Computational Semantics: An Introduction to Artificial Intelligence and Natural Language Comprehension*, North-Holland, Amsterdam.

L. WITTGENSTEIN, L.: *Tractatus Logico-Philosophicus*.

L. WITTGENSTEIN, L.: *Philosophical Investigations*.

L. WITTGENSTEIN, L.: *Zettel*, Basil Blackwell, Oxford, 1967.

L. WITTGENSTEIN, L.: *Remarks on the Philosophy of Psychology* (I), Basil Blackwell, Oxford, 1980.

L. WITTGENSTEIN, L.: *Remarks on the Philosophy of Psychology* (II), Basil Blackwell, Oxford, 1980.

L. WITTGENSTEIN, L.: *Last Writings on the Philosophy of Psychology* (I), Basil Blackwell, Oxford, 1982.