Original Paper UDC 130.11/Wittgenstein

Andrej Ule, Ljubljana

How Can One Catch a Thought-Bird?

Some Wittgensteinian Comments to Computational Modelling of Mind

Abstract

In this essay I analyse Wittgenstein's criticism of several assumptions that are crucial for a large part of cognitive science. These involve the concepts of computational processes in the brain which cause mental states and processes, the algorithmic processing of information in the brain (neural system), the brain as a machine, psychophysical parallelism, the thinking machine, as well as the confusion of rule following with behaviour in accordance with the rule. In my opinion, the theorists of cognitive science have not yet seriously considered Wittgenstein's criticism so they, quite surprisingly, frequently confuse the question »how does it work?« with »what does it do?« But their most »deleterious« mistake is their confusion of the internal computational (or parallel) processes taking place in the brain (which possibly cause mental states) with socially-based, everyday criteria of recognition and classification of, and knowledge about, the content of mental states.

The title of this essay is metaphorical. I believe that watching a human thought in action is as difficult as catching a flying bird (with one's hands). We have a chance to catch a bird only when it happens to come sufficiently close, perhaps pausing for a moment to sit on a branch next to us, or when it lands right in front of us to grab some food. But at that moment it is not a »flying bird«, but at most a bird that is just about to start flying (again). Yet, even in such moments our chances of getting hold of the bird are very small, a fact probably well known to anyone who has already tried to catch a bird. A human thought is much like a bird. Most of the time thoughts »fly« through our minds, meaning that they appear in our consciousness for a moment and then promptly disappear, i.e. are replaced with new thoughts. The very attempt to understand what happens inside our minds when we think some thought changes that thought into something else (another thought). In much the same way, we cannot understand how it is possible that with our thought we can 'touch' things that are entirely different from that thought, perhaps physically inaccessible or even non-existent. But as soon as we contemplate the thought in this manner, that thought 'escapes' us, and we end up staring in surprise, unable to understand what is happening. At times it may seem to us that the thought has stopped, so to say, that it stands clear in our sight and we will be able to seize it and understand what it is and what it means to »have a thought in one's mind« or to »think about this or that«. This phenomenon can be experienced when we, say, puzzle over a completely 'absorbing' problem and some particular thought is one of the assumptions that constitute that problem. However, as soon as we concentrate on that particular thought and move away from the problem which it constitutes, the thought often disappears promptly, i.e. escapes our mental horizon. This is one of the reasons why

assumptions about the problem are difficult to analyse – for us, they make sense only insofar as we experience them as part of the problem, but this sense is lost as soon as we analyse them as such.

Wittgenstein presents a similar difficulty in his *Philosophical Investigations* (*PI*, 1976).

»'The queer thing, thought' – but it does not strike us as queer when we are thinking. Thought does not strike us as mysterious while we are thinking, but only when we say, as it were retrospectively: 'How was that possible?' How was it possible for thought to deal with the very object *itself*? We feel as if by means of it we had caught reality in our net.« (*PI*, par. 428).

Thoughts seem to be an extremely elusive 'mental phenomenon'. We know what they are as long as no one asks us, but once asked, we no longer know, to paraphrase Augustine's witty remark about time. Many other 'mental phenomena' produce similar experiences, for example, understanding, intention, wish, will, feeling and so on.

In Wittgenstein's words, this is not the kind of question typically posed by the natural sciences, but

»… something that we know when no one asks us, but no longer know when we are supposed to give an account of it, is something we need to remind ourselves of. (And it is obviously something of which for some reason it is difficult to remind oneself.)« (*PI*, par. 89).

During his 'late stage' Wittgenstein pointed out how erroneous it is to seek causes and reasons for the elusiveness and indefiniteness of mental phenomena in their secretive nature, for example, strictly private 'internal' processes that are supposedly accessible only to pure introspection and about which one can speak only in the first person singular, with no other person being able to know about these processes except the person who is being introspective. According to this understanding, people can reach hypothetical conclusions about the mental states of others only on the basis of that person's visible behaviour, so the domain of the mental presumably fits into an entirely different order of things and events than the domain of the physical.

Wittgenstein strongly rejected this idea, arguing that it was absurd. Yet he also rejected the views of behaviourists, whom the elusiveness and subjectivity of mental states led to conclude that intellectuality, thinking, consciousness and the like were just appearances not grounded in reality, meaning that all we are left with is the physical behaviour of people. He speaks of the grammatical fiction of behaviourism (*PI*, par. 308) and says:

»How does the philosophical problem about mental process and states and about behaviourism arise? – The first step is the one that altogether escapes notice. We talk of processes and states and leave their nature undecided. Sometime perhaps we shall know more about them – we think. But that is just what commits us to a particular way of looking at the matter. For we have a definite concept of what it means to learn to know a process better… – And now the analogy which was to make us understand our thoughts falls to pieces. So we have to deny the yet uncomprehended process in yet unexplored medium. And now it looks as if we had denied mental processes. And naturally we don't want to deny them« (par. 309).

Wittgenstein points out that to talk about the thinking process, as a 'non-corporeal' process is misleading because thought cannot be separated from speech and behaviour. Such talk expresses our confusion arising from, on the one hand, our attempt to determine the meaning of the word 'to think' in a primitive manner, and on the other, our grammatical differentiation between, say, the grammar of the word 'to think' and that of 'to eat' (*PI*,

par. 339). He, however, adds that the difference in meanings seems to be too small (in much the same way as the difference between our saying that digits are real and numbers non-real is small).

This provides room for a 'third' possibility, i.e. an attempt to conceptualise intellect, thinking, experience and similar phenomena as 'internal' physical states and physical processes taking place within the physiological structure of an individual, say, within one's brain, as a result of certain internal mechanisms, e.g. computations taking place inside the brain. One advantage of this approach is that it preserves 'intuition' about the internal processes and states, a concept that so forcefully imposes itself on our consciousness, while at the same time rejecting the idea that these processes and states are inaccessible to scientific research. The notion of the internal physical processes and mechanisms in the human brain that cause mental phenomena or lead to them is undoubtedly very attractive for a large part of cognitive science. However, it has its principled limitations also pointed out by Wittgenstein. He opposed not just the idea that non-material processes taking place within an individual are inaccessible, but also the idea of the 'internal mechanism' or 'internal processes' in general. For him, the 'external–internal process' difference is simply too small to be useful in explaining the predicaments arising from everyday discourse and, even more so, philosophical discourse on mental phenomena.

According to Wittgenstein, of all the mental 'phenomena,' it is understanding and intention in particular that resist our attempts to conceptualise these as internal states or processes. In relation to these two phenomena, even the categories of 'state' and 'process' are questionable.

»The intention *with which* one acts does not 'accompany' the action any more than the thought 'accompanies' speech. Thought and intention are neither 'articulated' nor 'not articulated'; to be compared neither with a single note which sounds during the acting or speaking, nor with a tune. 'Talking' (whether out loud or silently) and 'thinking' are not concepts of the same kind; even thought they are in closest connection« (*PI*, II xi, p. 217).

Or, the same thought, expressed still more explicitly:

»Meaning it is not a process which accompanies a word. For no *process* could have the consequences of meaning. (Similarly, I think, it could be said: a calculation is not an experiment, for no experiment could have the peculiar consequences of a multiplication.«) (*PI*, II xi., p. 218).

The part in brackets is important because cognitive science often assumes that a calculation is a mental process equivalent to some physical process in a processor, or to a biophysical neural process. Indeed, such a process would more resemble a kind of 'micro-experiment' than a calculation, which is the idea strongly rejected by Wittgenstein.

In his *Remarks on the Philosophy of Psychology* (*RPP*, 1980), Wittgenstein even explicitly stated that no assumption seems more natural to him than the assumption that association, i.e. thinking, is not any kind of brain process, so the understanding of mental processes on the basis of brain processes is impossible (*RPP*, I., par. 903). As an illustration he uses the example of the seed. Identical seeds always produce identical plants, although *nothing* in that seed corresponds to the plant, so it is not possible to conclude which plant the seed will produce exclusively on the basis of its properties or structure. Only the history of the seed could possibly give us a clue as to which plant will grow from it. According to Wittgenstein, in a similar manner an entirely amorphous mass could produce an organism, almost with-

out reason. So, he wonders whether something similar could be the case with our thoughts, our speech or writing.[1] Later in the text Wittgenstein even goes so far as to raise doubt about the assumption that every mental process has a corresponding 'trace' in our brains, thus allowing the possibility that there exist psychological patterns that are not matched by *any* physiological patterns (*RPP*, I., par. 905). »If this turns upside down our notion of causality«, says Wittgenstein, »then it is time for it to be turned upside down«.[2] He also adds that the prejudice about the parallelism of the psychological and the physical is a result of our primitive interpretation of grammar. Because, if we allow that psychological phenomena follow a principle of causality that is not physiologically mediated, then we assume that there exists a kind of soul *besides* the body (*RPP*, I., par. 906).

This is not to say that Wittgenstein suddenly turned into an advocate of dualism, i.e. the thesis about a spiritual substance that exists apart from the body and enables spiritual phenomena. In fact, he was simply stressing, perhaps in a slightly exaggerated manner, a profound difference between human thinking and all other human processes or states, either material or non-material. In Wittgenstein's view, it is not possible to speak of thinking as a private process; it can be viewed only in the context of people's social activities. In other words, thinking takes place only within the context of rule following, within various linguistic games and within the interplay of human actions. These relationships cannot be translated back into the processes taking place within individuals, or into individual behaviours.

By saying that thinking is an internal process within our brain, i.e. a physiological process within the body that causes thinking or is thinking itself, we are in danger of cutting off thinking from the intersubjective practices of speech, mutual understanding, social functioning of people in various situations (contexts) and so on. Such a process would be, in a way, self-sufficient, and that would make it, after all, as 'private' as are the states and processes of the presumably spiritual substance. As long as we do not have available an intersubjectively verifiable (publicly accessible) and systematic link between assumed internal states and processes taking place within the brain, and between the individual's action in intersubjective (social) contexts, we are exposed to Wittgenstein's criticism of private speech, private states and processes. Of course, it is possible to hypothesize that such a link cannot be excluded in principle and that, at some point in the future, it might be possible to explicate. However, we are still left with an embarrassing gap between the publicly accessible, 'external' practice of people's actions in various situations (e.g. language games) that determines the meaning of our experiences, wishes, intentions, thoughts, and between the 'internal' (physical) nature, i.e. structure of these experiences, wishes and the like.

The point here is that this is not solely a conceptual rift between the substrates of mental states presumably entirely explicable by computation, but a rift between how 'processors' work, meaning processors that constitute the physiological-physical basis of mental states, and *what and how we are* when experiencing these states and expressing them in our social practice. The 'how we are' is essentially connected with the fact that our experiencing and functioning has *meaning for us*, that we *feel our existence* precisely through experiencing and functioning. The difference between the two is not just conceptual, but it is a difference in categories, and we have not the slightest clue as to how this rift could be overcome.

The very notion of the 'working' of a processor involves an ambiguity pertaining to description of the causal process, i.e. the sequence of the processor's states *vs*. the rules of processors' working. The former can be attained by listing the physical laws that regulate the physical flow of events and initial states of the processor at particular moments; to attain the latter, we could, for example, present the operation table featuring the processor's transitions from one state to another, depending on the input and the current state of the processor. Yet such a table would be a completely idealized creation, showing how the process *should* work in ideal circumstances rather than how it actually works. It is just one of the possible descriptions of the rules of working presumably implemented by the processor, not a description of the process. Such rules can also be presented in a table listing the operations that the processor can 'perform' provided those suitable inputs and suitable internal states are present. Yet whichever approach we adopt in determining the rules of processor's working, we have to keep in mind the difference between the idealized, *operational working* and the actual *flow* of events, i.e., the actual sequence of the processor's physical states. The operational working of the processor is the kind of working we *ascribe* to it when we *interpret* it as the implementation of a certain rule or tabular description of operation.

Cognitive sciences frequently refer to the 'Turing machine' model that is believed to represent the general formula for the computational working of natural or artificial processors. However, this model is just an idealized structure of the system's working to which is *ascribed* computational working, with no actual system corresponding to this model. The difference lies not just in the fact that a Turing machine uses an infinite memory tape, but also in the fact that this 'machine' is capable of returning, after several computational steps, to precisely the same state. By contrast, actual, finite automata never really 'return' to the same state, only to the 'same kind' of state that belongs in the same equivalence class of states with regard to a certain equivalence relation between the states. Therefore, it is a question whether in this case it is possible at all to speak about computation in the sense of 'operating with symbols'.

In his essay, that appeared in a collection of essays on the legacy of Turing, C. Fields explicitly stressed that it is necessary to distinguish between the behaviour of the physical system in time and the (dynamic) description of this system using an algorithm that »determines« changes in such a system based on (discreet) changes of a specific parameter (or parameters) that is observable (measurable) at the time of occurrence. The latter is *our* description method that enables us to map the changes of the system's states onto the algorithmic steps. In order to be able to give such a description, we must also have available certain interpretation that maps the system states onto the sequence of abstract computational states, whereby every a_k state of the system is 'obtained' by using the initial $_k$ steps of the algorithm at a specific input (Fields, 1996, p. 170). Fields points out that this kind of mapping is possible only if all operations that are listed in the process of mapping commutate algebraically (i.e. viewed from the perspective of quantum mechanics, measurements or observations of the process are mu-

1 Similar statements can be found in Wittgenstein's *Zettel* (*Z*) (1970, par. 608).

2 See also *Zettel*, par. 610.

tually independent). Fields refers to E. Ditriech, who in 1989 pointed out that such an interpretation can only be constructed if the system is antecendently, and informally, viewed as instantiating a function that can be computed by the algorithm, and if the *k* steps of the algorithm are recognized as computations of some antecedently understood subfunctions.

We need here a sufficient number of observed states of *the system* in order to map these onto the computational states traversed by the algorithm when it is applied to a given input. Fields speaks of such an interpretation of the system as a 'virtual machine' for the algorithm. However, this interpretation would be possible only if the behaviour of the system from an initial state (interpretable as an input to the algorithm) is always interpretable as an execution of the algorithm for the given input.

This is certainly an idealized assumption or, better still, an assumed idealization rather than the realistic flow of events within the system. The system, therefore, computes only inside the framework of our interpretation, as a virtual machine and not per se. A dynamic description of the system represents just a specific relation that gives a sufficiently accurate account of the sequence of (measurable, observable) physical states of the system in successive time intervals. In many cases it is possible to describe this sequence using a neat, continuous function, but the method is not essential – what is essential for the computational working of the system is an interpretation that presents the successive states of the system as a result of a specific algorithm (i.e. the system of operations) applied to the corresponding input states of the system that are the inputs for the virtual machine.

Fields' reflections lead him to two important conclusions. The first is

»… that any system that can be interpreted as simulating a virtual machine for some function *f* is a virtual machine for *f*. Similarly, any system that can be interpreted as simulating a universal computer is a universal computer. This feature of computation – that *simulating computing is computing* – is what sets it apart from an uninterpreted dynamic process such as fluid flow« (Fields, 1996, p. 171).

The second is

»… that a single physical system can often be interpreted as different virtual machines on the basis of different sets of measurements, by interpreting the values of different sets of variables as indicative of the state of the system« (ibid.).

Field's first thesis that perfect simulation of computation is computation should also be read *vice versa*, that is to say, every computation is its own 'perfect simulation', or such an interpretation of the working of the system that presents it as perfect computation (i.e., it presents every state of the system as a computing result of certain inputs and certain states of the system). As a result, there is no 'computation per se', i.e. there is any flow of events without a corresponding (computational) interpretation. This thought can be linked with Wittgenstein's interesting criticism (in *Tractatus*) of the equation of operation and predication, or operation and facts (events) (Wittgenstein, 1976). Wittgenstein actually argues that operations are an integral part of our symbolism, i.e. the manner of presenting the relations between the forms of the state of things rather than the states of things themselves, facts or events in the world.

In principle, we have to distinguish between operations and (prepositional) functions, which represent some relations in the world of facts. An operation represents formal relations, i.e. relations between forms rather than

between the individual representatives of forms. Wittgenstein illustrates this using the logical forms of sentences.

»A function cannot be its own argument, whereas operation can take one of its own results as its base« (*Tractatus*, 5.251).

»It is only in this way that the step from one term of a series of forms to another is possible« (5.252).

»One operation can counteract the effect of another. Operations can cancel one another« (5.253).

»Further an operation can also vanish (e. g. negation in the case of double negation)« (5.254).

In principle, we cannot expect functions (i.e. the descriptions of objects' properties or relations between them) to have the properties of operations. According to Wittgenstein, the 'internal' or 'formal' relations between the states of things, i.e. between the statements that describe these, cannot be expressed with sensible sentences but can only be presented by means of an operation that »represents a proposition as the result of an operation that produces it out of other propositions (which are the basis of the operation)« (5.21). Wittgenstein's assertion that operations are not functions (predications) also means that operations are not any state of things or facts, but they are our methods of representing the relations between the forms of the representatives of the state of things. In this sense, computations, too, if they are true operations represented by the Turing machine's schemes, do not occur in 'nature' but within the symbolic systems of representation.

Wittgenstein's indication of the fundamental difference between predication and operation additionally supports Field's distinction between computation and the flow of events in some physical system. To describe the flow of events, we use a sentence by which we relate the successive states of the system and put them into a specific order, perhaps relying on some mathematical function that expresses the natural law of occurrence. However, while such a description is not essential for the flow of events itself, for computation a certain (symbolic) description of the flow of events is essential. According to Wittgenstein, computation involves a successive (sometimes both successive and parallel) execution of operations that can be represented only by using some symbolic system, i.e. by using appropriate formal expressions, variables, operation tables or algorithmic procedures.

In his later works Wittgenstein drew attention to a similar distinction between the causal working of the machine and operational 'working' that occurs within the representation of a machine as a symbol. These reflections are especially relevant for discussion of the 'thinking machines,' i.e. any 'strong' program of artificial intelligence. Wittgenstein says that the machine can be conceptualised idealistically, meaning by abstracting possible breakdowns, malfunctions etc. In such a case it seems »that a machine has (possesses) such-and-such possibilities of movement« (*PI*, par. 193).[3] Once we know the machine, we imagine that everything else is determined in advance, say, a movement it will make. According to Wittgenstein, in such a case we actually use the machine or the image of the machine (or the diagram of the flow) as a symbol of a specific manner of functioning.

3
See also *Remarks on the Foundaments of Mathematics* (Wittgenstein, 1972, I. par. 125).

This is completely different from a prediction about the actual behaviour of a machine. Then we do not in general forget the possibility of a distortion of the parts and so on. The actual machine can always make a move completely different from the one 'predicted' by its symbolic counterpart, because in the symbolic counterpart the possibility of breakdown is *a priori,* i.e. grammatically, impossible. Wittgenstein hence concludes »that the movement of the machine-as-symbol is predetermined in a different sense from that in which the movement of any given actual machine is predetermined« (ibid.).

Yet even the talk about an 'ideal' or 'idealized' machine that is supposedly embodied by the machine-as-symbol misleads one into assuming that the machine has (possesses) its 'possibilities of movement.' Wittgenstein concludes that

»… this possibility of movement is not the movement, but it does not seem to be the mere physical conditions for moving either […] For while this is the empirical condition for movement, one could also imagine it to be otherwise« (ibid.).

He further says that to us it seems that an ideal possibility of movement is a kind of

»… shadow of the movement itself… We say: 'experience will show whether this gives the pin this possibility of movement' but we do not say 'experience will show whether this is the possibility of this movement« (*PI*, par. 194).

In talking like this, we erroneously interpret our own manner of talk, by taking the talk describing the operation of a machine-as-symbol to be a description of the actual working of the machine. Wittgenstein says that »we are like savages, primitive people, who hear the expressions of civilized men, put a false interpretation on them, and then draw the queerest conclusion from it« (ibid.). In short, a machine-as-symbol is a special *expression of the rules of working*, i.e. working according to the rule, rather than actual working. In a similar way a machine for conjunction represented by means of a corresponding table of the transition of states, would be a symbolic representation of a working by rule rather than the factual working of some 'logical machine', for example, a simple processor that implements conjunction.

Obviously, we have again arrived at the principled difference between operations within some symbolic system (model) and actual processes in the world. Regardless of how perfectly a machine works (for example, a machine that implements some logical operation), that still does not mean that that machine actually executes an operation, because even the most perfect machine can break down; its working may be interrupted, or it may be damaged, while in principle the execution of the operation excludes such a 'possibility'. For Wittgenstein, these possibilities are categorically, grammatically excluded.

However, Wittgenstein denied the possibility that any expression of the rule, i.e. rule following, represents the ultimate or original interpretation or explanation of rule following. In this sense he would not unconditionally agree with Field's formulation that computational interpretation of the system's working by a computational rule already represents the actual following of that computational rule, i.e. that the simulation of computation is computation. Actually, in so far as 'computational interpretation' is only a symbolic expression of the mapping of the system's physical states onto the

algorithmic states of the computation system, it is not possible to speak of computation. Only the execution of this mapping, that is, actual following of an algorithm is computation. In this sense, computation is just an actual computational interpretation rather than, for example, an 'imagined' formal interpretation that only describes the algorithm of computation. The table of mappings or the table of transitions is not yet an operation but a symbolic expression. If we combine the terminology used in *Tractatus* and *Philosophical Investigations*, we could say that actual computation only shows itself through the use of a specific algorithm in an environment such as is a social, publicly accessible practice (*PI*, par. 202), technique (par. 199) or custom (par. 199) in which we ourselves must participate logically, while to *describe* it we have to use some formal symbolism.

However, a sensible participation in the practice of rule following is not simply an 'adequate' response of the actor to certain inputs, but participation in the life form of beings for whom rule following is a sensible and basically unquestionable, understandable action. This non-questionability must simply be accepted as a basic fact.

»What has to be accepted, the given, is – so one could say – forms of life« (*PI*, II xi, p. 226).

The shared human form of life is not something that is fixed, but our common way of agency. It is a net of activities that determines our understanding and agency. Wittgenstein speaks about this in *Zettel* (*Z*, par. 576).

»How can we describe the human way of action? Well, only by showing how the actions of various people swarm in all directions. It is not what one thinks right now, an individual action, but the whole swarming of human actions that forms the background against which we observe every particular action and which determines our judgment, our notions and reactions.«

For Wittgenstein, the rules, i.e. rule following as well as operations and execution of operations, are an inseparable component of the human form of life, i.e. »the swarming of human actions«. For this reason, rule following cannot be abstracted and observed separately from this background – for example, in such a way as to ask whether some machine or organism in itself follows the rule.

Since thinking implies the ability to follow rules, Wittgenstein argues that it is not sensible to ask whether, for example, the machine thinks. Or, to be more precise, such a question would imply a certain resemblance with the human being.

»Could a machine think? Could it be in pain? – Well, is the human body to be called such a machine? It surely comes as close as possible to being such a machine (*PI*, par. 359). But a machine surely cannot think! Is that an empirical statement? No. We only say of a human being and what is like one that it thinks. We also say it of dolls and no doubt of spirits too. Look at the word 'to think' as a tool.« (*PI*, par. 360).

Wittgenstein, therefore, does not reject completely the possibility that a machine could think, but only under the assumption of a certain resemblance with humans, if only a virtual one. But whence could such a resemblance arise? Of course, from our representation that the working of a machine is part of the human form of life, that it is a logical part of the net of human agency, for example, the net of human communication, cooperation among people in common projects, games etc. All that we can say is that, if we could imagine a robot that accurately imitates human actions in paradigmatic life situations and participates on an equal footing in human communication, then we could imagine such a robot as a 'thinking one'. This is

similar to how we imagine the behaviour of 'live puppets' (e.g. Pinocchio) or spirits. Yet, the road from imagination to implementation is long, perhaps even impossible.

In the last sentence in the quotation given above, Wittgenstein hints that this is as if thinking were a tool. Can thinking be a tool? Perhaps in a manner similar to that in which our hands are a 'tool' for work. Undoubtedly, this is one possible perspective on thinking, but it involves just one aspect and fails to take account of the whole. We can view thinking as if it were a tool (created by evolution, nature, God etc.), but the very fact that we view thinking as if it were a tool says that thinking in itself is not a tool. For example, we cannot view humans 'as alive' or 'as thinking', but we simply hold humans to be alive and thinking. We could view a human being »as if it were a puppet«, but that is not its 'nature'. Similarly, we could view a puppet not »*as if it were* a simulation of a man« but as *being* a simulation of a man (for us). And I could view the puppet as if it was alive, a thinking being, but that is not its 'nature'.

Does Wittgenstein's rejection of the mental phenomena seen as internal states within the individual's neural and physiological structure mean that he abandoned scientific psychology? Does it mean that we should interpret psychological phenomena only as social phenomena, or at least 'interpersonal' rather than 'personal'? Indeed, Wittgenstein is quite frequently viewed as an opponent of scientific psychology (see e.g. Williams, 1999, Rey, 2003), but his piercing remarks about philosophy would equally, or even more strongly, 'justify' the assertion that he was an opponent of philosophy as well. In my opinion such conclusions are premature. It is more probable that Wittgenstein strived to 'get rid' of all psychological, philosophical and other kinds of theories about human intellectuality, speech, communication and so on, because he attempted to expose that which in his opinion was no theory but the manifestation of the 'firm ground', i.e. unambiguous mastery of rules and language. Since some psychological and philosophical theories attempted to encompass this 'ground' and explained it by using natural scientific, psychological and philosophical assumptions and theories, they only further aggravated perspicuous presentation (*PI*, par. 109, 122), and Wittgenstein fervently resisted this. In his *Remarks on the Philosophy and Psychology*, he even outlined a kind of plan for the analysis of psychological concepts, but everyday concepts rather than those invented by science for its own purpose (*RPP*, II, par. 63, 148). This means that when creating psychological concepts we must not break the link with their everyday understanding. This is a trait that distinguishes psychology from some natural sciences that pursue the greatest possible independence from the everyday understanding of their subject areas. The preservation of the systemic link between the everyday understanding of psychological concepts and phenomena does not imply elimination of scientific psychology or its reduction to behaviourism. As Mary McGrinn says, Wittgenstein's criticism of psychological concepts does not mean that psychological concepts require behavioural criteria of application but is »an attempt to show that we cannot derive an idea of what a given psychological state is simply through introspection« (McGinn, 1997, p. 130). According to McGinn, the moral of the argument against private language

»… is not that our psychological concepts *must* possess public criteria, but that it is only by reminding ourselves of the grammar of our ordinary psychological concepts that we can grasp the essence, or nature, of a given kind of psychological state« (ibid.).

I agree with this interpretation.

Undoubtedly such an understanding of psychology and psychological concepts quite opposes the intentions of a large part of cognitive science, and particularly those of the radical representatives of computationalism. So, in my opinion, Wittgenstein cannot be described as a forerunner of cognitive science, or of the computer metaphor or the like (although he was a good friend of Turing).

I will take a look at two such interpretations: G. Rey's essay *Why Wittgenstein Ought to Have Been a Computationalist* (2003) and J. Leiber's book *An Invitation to Cognitive Science* (1991). In Rey's view, what separates Wittgenstein from cognitive science is mainly his excessive preoccupation with the first/third person problem. According to Wittgenstein, psychological concepts are based on divergences between the first and third person (Rey, 2003, p. 240). Wittgenstein abandons a theory of meaning that presumes that words enjoy a uniform referential relation to objects. Rey has pointed that Wittgenstein sketched instead a theory involving a term's role in a context, language game, or form of life (p. 241). According to that theory, »mental predicates like 'hopes', 'expects', 'ardently loves' are much more *widely relational* than the traditional conception of 'inner processes' allows« (ibid.) Rey correctly pointed to the fact that even Wittgenstein's use of 'behaviour' itself is evidently intended to be thus broad. Rey discussed three sources of support for Wittgenstein's suggestions on the relational nature of mental predicates and human behaviour in the modern philosophy of mind: externalist intuitions about content, verificationist and 'criteriological' theories of meaning, and functionalist conceptions of mental states. Rey, however, tries to support all of these theories, especially the functionalist approaches to meaning and mental states with some 'mentalist' hypotheses. The first of these is *Modest Mentalism* (*MM*), which proposes at least two basic kinds of mental states, informational ones that represent the world, and directional ones that direct their agent towards or away from some represented state (p. 245). The second is the *Causal/Computational-Representational Theory of Thought* (*CRTT*), which states

»… that we should regard the brain as a computer performing operations in real time on logically complex internal representations, the primitives of which stand in certain co-variational relations with either stimulus pattern or with phenomena in the environment« (p. 249).

The third hypothesis is the *completion of CRTT* with *some internal mechanism* which produces *qualia* experiences, e.g. experiences of colours, tastes of food, feeling of pain, etc. Ray claims

»… that such experiences are states involving a particular computational relation to specially restricted predicates in a creature's system of internal representation« (p. 258).

Rey believes that mental states supervene on states of the agent's brain and that the appropriate computational mechanisms can give us some additional 'outward criteria' of mental states[4] which some purely behavioural criteria cannot give us (p. 257, 259). He can save talk of literal inner processes without the assumption of their principial privacy. This theory »needs some story with the kind of psychologically plausible details it provides« (p.

4
Rey refers to the (in)famous Wittgenstein's remark that »an inner process stands in need of outward criteria« (*PI*, par. 580).

262), which was absent in Wittgenstein. Rey further »believes that he doubts it is only philosophers who are captivated by the naïve, introspective picture of the mind, or only language is its source« (ibid). He is probably correct here. The naïve picture of mind or language is a set of simple analogies or models taken from nature, and projected onto man. It is true that »ordinary language can sustain and dispense with any number of good and bad analogies; people can get captivated by bad analogies if there isn't a more sensible one to replace it« (p. 262). However, we have to be aware of the suggestive 'power' of these analogies in order to prevent their non-reflective use in science or philosophy. Additionally, Rey claims that such naïve pictures are mostly our largely involuntary reactions to things that look and act like our conspecifices. We project these pictures into them correlative to that reaction in ourselves and are, indeed, unwilling to project them into things that do not induce that reaction (p. 264).

In contrast to Rey, I think that the very idea of 'inner processes' in our brains which produce mental states is one such powerful analogy that postulates some provisory outward criteria for mental processes and their causal explanation but it still misleads us. I have tried to show that Wittgenstein opposes the very concept of the inner process, either non-material or material. We can extend Wittgenstein's remark that it is

»… meaning not a process which accompanies a word. For no *process* could have the consequences of meaning« (*PI*, II xi., p. 218)

to intending, seeing aspects, calculation, etc., and therefore, to many kinds of thinking. These are Wittgenstein's principal statements which cannot be overridden or 'improved' by some more 'scientific' picture of inner processes. Wittgenstein's criticism of the machine metaphor clearly puts the principal difference between *operations* in the *symbolic* model of the machine working, and possible real (mis)working of the machine. It isn't only the difference between the machine, taken in idealisation, in abstraction from possible malfunctioning, and real working of the machine, but the categorical difference between rule following in the symbolic model, and sequences of events (states). Rey does briefly refer to the rule following problem, and agrees with the »normativistic« concept of rule following according to which »rule can be 'normative' insofar as a particular mechanically realized algorithm might be the best idealized explanation of the actual operation in the brain« (Rey, 2003, p. 253 (rem. 48)). Rey agrees with the opinion of Howich, Pietroski, et al. that the Kripke's famous paradox on rule following

»… has to do with the resources of idealization […], and not about normativity of rules-or the presence of 'surroundings that make standards into standards' -but about whether it is reasonable to construe the nervous system as realizing a particular mental competence« (ibid.).

However, if an algorithm may be the best idealized explanation of the actual operation in the brain, then it is *not* the real (actual) brain process that is performing an operation, or follows a rule but *we* who explain this process by a 'normative' algorithm. It is quite unclear to speak on »mechanical realized algorithm in the brain« because this 'realization' isn't an algorithm. The brain 'works' algorithmically only through our (active) interpretation as a computing process (see the discussion of Field's theses), not 'per se'.[5]

In *An Invitation to Cognitive Science* (1991) J. Leiber argues that Wittgenstein is connected to cognitive science primarily through his rejection of the 'old paradigm' of psychology, i.e. the dualistic theory of an embodied

soul and the principle that we just know, by introspection, and know in their entirety our mental states (p. 46, 70). The demolition of the old paradigm leaves the way open for the 'new paradigm', that is the computational theory of the mind (p. 65). Leiber believes that Wittgenstein is a 'cognitive naturalist' (p. 61, 159). Leiber cites language-learning, face recognition, aspect-perception and rule-following as some important common research projects that connect Wittgenstein to cognitive science. Leiber strongly criticized the view of Wittgenstein as an opponent of scientific psychology. I will consider only one point in Leiber book, the claim that Wittgenstein's account of meaning and understanding leads naturally to a computational view of mind (p. 67–68), and that he anticipated the computational model of mind (p. 109). Leiber refers to Wittgenstein's thesis that the meaning of an expression is its role in a language-game. He interprets this thesis as the thesis that the meaning of an expression is its place in a formalised procedure (p. 77–78). Leiber refers here to Wittgenstein's example of the language-game of a shopkeeper in *Philosophical Investigation*.

Someone requested from the grocer five red apples. The grocer first opens the drawer marked 'apples', then he looks up the word 'red' in a table and finds a colour sample opposite it; then he says the series of cardinal numbers – up to the word 'five' and for each number he takes an apple of the same colour as the sample out of the drawer. It is in this and similar ways that one operates with words (*PI*, par. 1). Leiber interprets this language-game so that the meanings of the words 'five', 'red' and 'apple' are just their roles in 'step by step physical procedures' (p. 66). For Leiber, the grocer's understanding of the utterance 'five red apples' consists in his mastery of a technique, namely the simple competences of fetching and matching. Leiber believes that a computer can do the same procedure as the grocer performed. Thus he concludes with the thesis that Wittgenstein's account of meaning and understanding anticipated the computational model of the mind.

We have seen that Wittgenstein strongly opposed the computational model of the mind (and the representational hypothesis of thinking and speaking too). Leiber has first to show that the step by step procedures involved in the grocer's language-game are in fact algorithmic and can be carried out by computer, and second, that other language-games occur in the same way, that is, as performing some algorithmic procedures which can also be performed by computers.

First, it is hard to see that the grocer's procedure in the language game is in fact algorithmic because the grocer's actions can be much more complicated than they seem at first glance. It could be that the complete description of the physical procedure transcends the possibility of any algorithm, and thus they cannot be simulated on a computer. Sure, we can interpret the given language-game in an *idealized* model but it would be only a formal *symbol* of the language-game where some algorithmically defined operations can be »done«, and not the real process of the language-game. Sure, our every-day understanding of this language-game as a step by step process can be very close to the algorithmic model of the game, but this understanding needn't be very close to the pure physical process. For exam-

5
I give more intensive analysis of the rule following problem, and the Wittgenstein's evolution of the concept of operation into the concept of rule-following in my book *Operationen und Regeln bei Wittgenstein* (1997).

ple, our intuitive »parsing« of the grocer's behaviour in the language-game could be very different from the real »elements« of his behaviour which are modelled by a computer, because we usually follow the language description of the game, and not the adequate physical description of the events.

Leiber also didn't present any proof of the thesis that all other language-games could be present algorithmically, as in the grocer's case. He refers to Wittgenstein's claim that language-games are part of our natural history, and claims that we have a natural history of formalizing (p. 68). In the Leiber's account, *all* language-games involve formalised procedures but some independent evidence does not support this thesis. The same objection that was raised to the grocer's language-game can apply to all of them. The human ability to perform step-by step procedures can be, in fact, the result of our natural history, but these procedures and our competence for executing such procedures could still be much too complicated for any exact formal description by algorithms which could be performed by a computer.

There is another important question that is similar to that I raised in the section discussing Rey's thesis. This question is whether Leiber truly succeeded in showing that algorithm-based processes in the brain can explain our ability to follow rules and whether the computer simulation of these processes corresponds to human rule following. A computer can indeed »behave in accordance with the rule« yet not actually follow the rule, because it lacks the required competence. For Wittgenstein, this competence develops within the community which pursues the practice of rule following, adheres to certain customs, trains rule following and various techniques of rule following, etc. Lieber would have to show how this activity could be translated into algorithmic processes that could be potentially simulated by the computer, if he wanted to show that all language games contain formally describable algorithmic processes. Alternately, he should refute the distinction Wittgenstein made between behaviour »in accordance with the rule« and competitive knowledge of the rule.

D. Proudfoot and B. J. Copeland have recently presented a comprehensive criticism of Leiber's book, which I myself have followed, although their criticism is more extensive and includes, for example, the criticism of representational understanding of meaning, speech and perception in the computational models of cognitive science. Proudfoot and B. J. Copeland reject Leiber's conviction that Wittgenstein aimed at a scientific understanding of the mind that transcends everyday understanding, and that Wittgenstein's explanation of cognition was causal in its intention, i.e. proto-scientific (Proudfoot, Copeland, 1994). The authors point out the difference between the conceptual and causal explanations that is similar to the difference between the 'what' and 'how' questions used by many cognitive scientists (for example by N. Block). In his criticism of functionalism, Block argues that the question 'How does the processor work?' is not *a question for cognitive science to answer*. This question may belong in another discipline, electronic circuit theory. However, we must distinguish the question of *how something works* from the question of *what it does*. For Block, the question of *what* a primitive processor does is a part of cognitive science, but the question of *how* it does is not (Block, 1990, p. 257). Proudfoot and Copeland claim that the »primitive component – the 'endpoint – in the explanation of, e.g., meaning and understanding, is the natural behaviour of the human being at the level of ordinary lived experience«. However, clari-

fying the role of natural human behaviour in creating the forms of life and language-games generating our ordinary concept of mind is a matter for philosophy. The investigation of the causal mechanism underlying such behaviour is not (Proudfoot, Copeland, 1994, p. 514–515). They claim that »cognitive science, in so far as it is the further investigation of what for Wittgenstein is a primitive component, is nothing more than a 'realization science'«. Assuredly, this role doesn't suit the current conceptions of many philosophers of the mind. Thus, they conclude that, »contrary to Leiber's view, Wittgenstein is no cognitive scientist, even one in heavy disguise«, and he was *neither* a mentalist nor a behaviourist, *neither* a proponent of Leiber's 'old paradigm' nor a cognitive scientist. The old and the new paradigms do not present a genuine dichotomy, as many people assume (ibid.).

These thoughts appropriately conclude my reflections on how to capture thought in the web of cognitive science. In my opinion, the theorists of cognitive science have not yet seriously considered Wittgenstein's criticism of »internal processes« (and states), and surprisingly they frequently confuse the question »how does it work?« with the question »what does it do?«. Similarly, they confuse causal explanation with operations in a symbolic model (or, more generally, with rule following), as well as behaviour in accordance with the rule with rule following. But the most »serious error« is the confusion of internal computational (or parallel) processes in the brain that may cause mental states with external, socially, based everyday criteria of recognition, classification and knowledge of the content of mental states.

References:

Block, N. (1990): »The computer model of the mind«. In D. N. Osherin, H. Lesnik (eds.), *An Invitation to Cognitive Science. Vol. 3. Thinking*. Cambridge/MA: MIT, 247–289.

Dietrich, E. (1989): »Semantics and the Computational Paradigm in Cognitive Psychology«. *Synthese*, 79: 119–41.

Fields, C. (1996): »Measurement and Computational Description«. In: P.J.R. Millican, A. Clark, *Machines and Thought, vol. 1*, Oxford: Clarendon Press, 165–177.

Leiber, J. (1991): *An Invitation to Cognitive Science.* Oxford B: Blackwell.

McGinn, M. (1997): *Wittgenstein and the Philosophical Investigations*. London, New York: Routledge.

Proudfoot, D., Copeland, B. J. (2004): »Turing, Wittgenstein and the Sciences of the Mind. A Critical Notice of Justin Leiber 'An Invitation to Cognitive Science'«. *Australasian Journal of Philosophy*, vol. 74.

Rey, G. (2003): »Why Wittgenstein Ought to Have Been a Computationalist?«. *Croatian Journal of Philosophy*, Vol. III, No. 9: 231–264.

Ule, A. (1997): *Operationen und Regeln bei Wittgenstein.* Frankfurt/M.: P. Lang Verlag.

Wittgenstein, L. (1972): *Remarks on the Foundaments of Mathematics*, Cambridge/MA: MIT Press.

Wiitgenstein, L. (1970): *Zettel*. In L. Wittgenstein, *Schriften 5*, Frankfurt/M.: Suhrkamp.

Wittgenstein, L. (1974): *Tractatus logico-philosophicus (Tr)*. London: Routledge & Kegan Paul, Wittgenstein, L. (1976): *Philosophical Investigations (PI)*. Oxford: Blackwell.

Wittgenstein, L. (1980): *Bemerkungen über Philosophie der Psyhologie (Remarks on the Philosophy of Psychology)*. Frankfurt/M.: Suhrkamp.

Andrej Ule

Wie kann man einen Gedankenvogel fangen?

Einige Kommentare von Wittgenstein zur komputationalen Formung des Geistes

In diesem Essay analysiert der Autor Wittgensteins Kritik an einigen Annahmen, die für einen Grossteil der Kognitionswissenschaft von zentraler Bedeutung sind. Diese umfassen die Konzepte von komputationalen Prozessen im Gehirn, die mentale Zustände und Prozesse hervorbringen, die algorythmische Informationsprozessierung im Gehirn (neurales System), das Gehirn als Maschine, den psychophysischen Parallelismus, die Denkmaschine sowie die Konfusion der Regel, die dem Benehmen folgt im Einklang mit dieser Regel. Nach Meinung des Autors haben die Theoretiker der Kognitionswissenschaft Wittgensteins Kritik noch immer nicht ernsthaft erörtert, so dass sie, was verwundern mag, häufig die Frage »Wie funktioniert das?« mit der Frage »Was macht das?« verwechseln. Doch ihr »verhängnisvollster« Fehler besteht in der Verwechslung interner komputationaler (oder paralleler) Prozesse, die im Gehirn stattfinden (und die möglicherweise mentale Zustände erzeugen) mit sozialbegründeten, alltäglichen Kriterien des Erkennens und der Klassifizierung des Inhalts und des Wissens vom Inhalt mentaler Zustände.

Andrej Ule

Comment peut-on attraper l'oiseau de la pensée?

Les commentaires certains du models computationnels de Wittgenstein

Dans cet essai, j'analyse la critique que Wittgestein fait d'un certain nombre de thèses qui sont cruciales pour une grande partie de la science cognitive. Il s'agit notamment des concepts de processus computationnels dans le cerveau qui causent des états mentaux, du traitement algorithmique des informations dans le cerveau (le système neuronal), du cerveau comme machine, du parallélisme psychophysique, de la machine pensante, ainsi que de la confusion du fait de suivre des règles avec le comportement qui est en conformité avec ces règles. À mon avis, les théoriciens de la science cognitive n'ont pas encore examiné sérieusement la critique de Wittgenstein, de sorte que, chose étonnante, ils confondent souvent la question »comment cela fonctionne?« avec la question »qu'est-ce que cela fait-il?«. Mais leur erreur »capitale«, c'est de confondre les processus computationnels internes (ou parallèles) se déroulant dans le cerveau (qui peut-être causent différents états mentaux) avec les critères quotidiens, socialement basés, de reconnaissance, de classifications et de connaissance des contenus des états mentaux.