University of Reading

Wittgenstein and his legacy

Book or Report Section

Accepted Version

Schroeder, S. ORCID: https://orcid.org/0000-0002-4480-6458 (2018) Wittgenstein and his legacy. In: Kind, A. (ed.) Philosophy of Mind in the Twentieth and Twenty-First Centuries. The History of the Philosophy of Mind (6). Routledge, London. ISBN 9781138243972 Available at https://centaur.reading.ac.uk/39625/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

Published version at: https://www.routledge.com/Philosophy-of-Mind-in-the-Twentieth-and-Twenty-First-Centuries-The-History/Kind/p/book/9781138243972

Publisher: Routledge

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

University of Reading

Reading's research outputs online

Wittgenstein and His Legacy

Severin Schroeder (University of Reading)

After a brief account of Wittgenstein's conception of philosophy (1.), I shall describe what I call the 'inner-object model' of mental occurrences, i.e. the kind of dualist position that Wittgenstein argued against in different areas of the philosophy of mind (2.). The following sections present his discussion of this prevalent misconception with regard to bodily sensations and other minds (3.), understanding (4.), thinking (5.), and voluntary action (6.). Finally, I shall briefly consider the relation between Wittgenstein's views and functionalism (7.).

1. Wittgenstein's conception of philosophy

As a young man Wittgenstein had made himself a name as one of the foremost philosophers of his age by the publication of his *Tractatus Logico-Philosophicus* (1921). In this first book, under the influence of the founding fathers of modern formal logic, Gottlob Frege and Bertrand Russell, he offered a unified theory of logic, language, and the metaphysical structure of the world. However, when some ten years later, after having worked for a while as a primary school teacher and then as an architect, he returned to philosophy, Wittgenstein found his earlier views deeply flawed. He began to rethink not only the details of his earlier doctrines, but the whole approach: the whole conception of philosophy underlying them. The result was his second book, the *Philosophical Investigations*, published only after his death. In it he first offers a devastating critique of his earlier theories (not of their technical details, but of their underlying assumptions), then sketches his startlingly new approach to philosophy, and finally, applies this new approach to various problems in the philosophy of mind.

In a famous polemical passage in his *Enquiry* Hume claimed that an academic discipline that was neither mathematics nor empirical science could only result in 'sophistry and illusion' (Hume 1748, 165). How can philosophy, as a non-

mathematical *a priori* discipline avoid this verdict? Wittgenstein's response to Hume's challenge is that philosophy is indeed concerned with sophistry and illusion. However, what Hume failed to realise is that philosophers do not only *produce* sophisms and illusions, but that there is also the perfectly respectable, critical philosophical activity of *dissolving* sophisms and *dispelling* illusions; a philosophy that seeks clarity and understanding of difficult conceptual relations, rather than new knowledge or theories.

Philosophical problems come about when we fail to understand, or are even confused by, our concepts and the ways they relate to each other. We tend to have an overly simplified picture of the workings of language which when applied to certain concepts is likely to result in paradox. For example, we naturally expect nouns to be name of (kinds of) objects (*BB* 1), as suggested by words like 'tree', 'house' or 'foot'. Therefore, when trying to understand the meaning of a word such as 'mind', we naively assume that it too must stand for a kind of object or substance, the question being only what kind of object that could be: a material object (say, the brain) or perhaps a mysterious non-spatial soul substance (as envisaged by Descartes). Both alternatives lead to implausible results. The mistake was to construe the word 'mind' as a name of a thing in the first place (cf. *PI* §308). Hence, the very question 'What (kind of thing) is the mind?' should not be answered, but rejected, as it is based on a conceptual misunderstanding. Frequently, a philosophical problem is a confusion 'expressed in the form of a question that doesn't acknowledge the confusion' (*PG* 193).

In order to dissolve such philosophical problems we need to pay more careful and unprejudiced attention to the actual functioning of the concepts in question. Hence we assemble reminders of common usage (and common sense) (*PI* §127), and we give synoptic representations of the use of the words involved (*PI* §122). The result of such philosophical investigations will be *clarity*: the demolition of illusions (*PI* §118), a better understanding of the relations between certain concepts. It will not, however, lead to any new and surprising insights or theories (*BT* 419). If one wanted to characterise Wittgenstein's philosophy in one sentence, one could say that

it is a defence of common sense against some clever forms of nonsense (or patent falsehood) suggested by a misunderstanding of the forms of our language.[1]

2. The inner-object model

The starting point of Wittgenstein's later philosophy, and one of its leitmotivs, is a critique of his own earlier ideas about language, in particular of a view that could be called *referentialism*: the view (already illustrated in the previous section) that 'a word has meaning by referring to something' (*PO* 454), and that 'language consists in naming objects, namely: people, species, colours, pains, moods, numbers, etc.' (*BT* 209v). Later in the *Philosophical Investigations,* having discussed and rejected referentialism in general, Wittgenstein considers various specific instances of referentialism, and further confusions and problems to which it gives rise, in the philosophy of mind. Referentialism applied to psychological terms — such as 'sensation', 'thought', 'understanding', 'willing' — results in a philosophical picture which can be called *the inner-object model*. If words (at least nouns and verbs) stand for objects, then psychological words must presumably stand for inner objects, perceived by one's inner sense inside one's mind, which we tend to think of as a private container to which only its owner has access. (Of course the word 'object' must be taken in the broadest sense to cover any objective phenomena, including states, processes and events.)

'The inner-object model' is basically another name for 'Cartesian dualism'. But whereas the latter label emphasises Descartes's distinction between the mental and the physical, the former indicates the surprising thrust of Wittgenstein's critique: that on Descartes's picture the difference between the mental and the physical, far from being overly pronounced, appears rather too slight! The psychological realm is construed in parallel to the physical realm: thoughts and feelings are regarded as objects like chairs and tables — only located in a private mental space rather than the public physical space. Against this picture Wittgenstein is going to argue that the

[1] For further discussion and illustration of Wittgenstein's conception of philosophy, see Schroeder 2006, 151-68.

differences between psychological and physical concepts are far greater than commonly assumed.

3. Sensations and other minds

In §243 of the *Philosophical Investigations* Wittgenstein presents the idea of a private sensation language: 'The individual words of this language are to refer to what can only be known to the person speaking; to his immediate private sensations. So another person cannot understand the language' (*PI* §243). This passage naively expresses the dualist, or inner-object view (and not, as the following discussion shows, Wittgenstein's own position): It is simply assumed that sensations, feelings, moods and the rest are private, inner objects, inaccessible to others. Wittgenstein's procedure in §§243-315 of the *Investigations* is to develop the consequences of that view with respect to words for bodily sensations and feelings and then to show how those consequences lead to absurdity or contradiction. Chief among those consequences is the following: as an inner object a bodily sensation — and indeed the mind on the whole — is logically independent of any behavioural manifestations; just as the contents of a box are logically independent of the label on the box. From this follows the problem of other minds: If minds are logically independent of behaviour, how can we ever know for certain what others think or feel, or indeed, whether they think or feel anything at all? There is always the possibility of deception: people can hide their feelings and simulate feelings they do not have. And there is also the deeper worry that the contents of our minds may be, to some extent, incommunicable. How do you know whether what you call your 'pain' is at all like the private experience I call 'pain'? When I give names to my feelings the meanings of these names are, strictly speaking, as inaccessible to you as those private feelings of mine. You may guess what I feel, and hence what my words mean, but you can never be certain about it. Thus the inner-object model of sensations leads to the idea of a strictly private language, one that could not possibly be understood by anybody else.

I shall now reconstruct Wittgenstein's principal objections to the inner-object view of sensations.

(i) A first objection can be called the *idle-wheel argument*. Suppose that when people complain about pains they have experiences that vary dramatically from person to person, but are so insuperably private that these differences can never be ascertained. Then, these differences can never affect the public use of the word 'pain', for example the way one talks about one's ailments to one's doctor. Hence, where the meaning of our public word 'pain' is concerned, any entirely private occurrence that might accompany the use of that word 'drops out of consideration as irrelevant' (*PI* §293). As Wittgenstein puts it in §271: 'a wheel that can be turned though nothing else moves with it, is not part of the mechanism'.

(ii) Another line of attack against the inner-object model of sensations concerns the Cartesian idea of a privileged knowledge of the contents of one's own mind:

> In what sense are my sensations *private*? — Well, only I can know whether I am really in pain; another person can only surmise it. — In one way this is false, and in another nonsense. If we are using the word "to know" as it is normally used (and how else are we to use it!), then other people very often know when I'm in pain. — Yes, but all the same not with the certainty with which I know it myself! — It can't be said of me at all (except perhaps as a joke) that I *know* I'm in pain. [*PI* §246]

Of course, by ordinary standards, we often do know when others are in pain. In order to make scepticism about other minds appear at all plausible some more demanding standard needs to be invoked, and that is what seems to be the infallible knowledge one has of one's own sensations. By comparison with this paradigm of 'real knowledge' it would indeed appear that not much else can be 'known'. Wittgenstein's reply swiftly turns the tables: Far from being a paradigm of knowledge this is not really a case of knowledge at all! Why not? The crucial point is that what we ordinarily call knowledge presupposes the logical possibility of error and ignorance. You can be said to know something only where it would also have been conceivable for you not to know it. Just as you cannot meaningfully be said to be the winner of a game in which nobody can lose, there is no sense in speaking of knowledge where there is, logically, no possibility of ignorance, doubt or error (*PI* p.221: PPF §311). Hence, since one cannot be mistaken or in doubt about one's own sensations (*PI* §288), one cannot really say that one knows of one's own sensations

either. Thus scepticism about other minds is stopped in its tracks: By ordinary standards it is undeniably possible often to know what others feel. And this cannot be said to be only an inferior kind of knowledge compared with my knowledge of my own pain, for I do not have *knowledge* of my own pain. The ability to express one's own feelings is not correctly described as knowledge.[2]

(iii) There is yet another objection to the view that one cannot know but only surmise what others feel. Although apparently a consequence of the inner-object assumption, it is in fact inconsistent with that assumption. The claim that I know sensations only from inner experience is incompatible with my attributing sensations, though precariously, to others. It does not even make sense to assume that something like one's private experience *might* also be had by somebody else. This objection, which goes to the core of the inner-object model, may be called the *ascribability argument*. It is tempting to think: 'If I suppose that someone has a pain, say, then I am simply supposing that he has just the same as *this*' (cf. *PI* §350). Wittgenstein's reply is that the proponent of the inner-object model lacks the conceptual resources for this transition from introspection to talk about the feelings of others:

> If one has to imagine someone else's pain on the model of one's own, this is none too easy a thing to do: for I have to imagine pain which I *don't feel* on the model of the pain which I *do feel*. That is, what I have to do is not simply to make a transition in the imagination from pain in one place to pain in another. As, from pain in the hand to pain in the arm. For it is not as if I had to imagine that I feel pain in some part of his body (which would also be possible). [*PI* §302]

According to the inner-object view, one learns what pain is by *having* pain. When, for example, I have hurt myself, I concentrate on the feeling and impress it upon myself that *this* is what one calls 'pain'. But, Wittgenstein objects, even *if* that procedure allowed me henceforth to identify my own pains correctly, it would not enable me to understand statements of the form '*NN* is in pain'. Introspection can never teach me how to ascribe a sensation to a particular person — not even to myself. For when I feel pain, I do not feel *my self* having the pain. I only feel pain in

---

[2] Admittedly, it is possible to use the expression 'I know I am in pain' in a meaningful way; but then the word 'know' functions rather differently from its ordinary employment. It may, for example, be used in a bad-tempered request not to be told what one doesn't need to be told.

a certain place, so the idea that others have pain I could only understand to mean that I feel pain in other bodies.

The difficulty is not how to make the transition from 'I am in pain' to 'He is in pain', but from 'There is pain' to 'He is in pain'. From the point of view of introspective consciousness that is acquainted with pain only as something *felt*, the idea of another person's pain amounts to pain-that-is-not-felt — which must appear as a contradiction in terms. Roughly speaking, feeling pain (my own pain) cannot teach me to understand the idea of pain that is not felt (others' pain).

In fact, in order to make sense of the assumption that someone else is in pain it is not the *experience* of pain I need, but the *concept* of pain. And a grasp of this concept includes an understanding of what it means to ascribe pain to a particular person. I must know, in brief, that 'the subject of pain is the person who gives it expression' (*PI* §302). Of course, a pain need not be expressed: one can in many cases keep one's sensations to oneself. Moreover, an expression of pain may be faked. Still, it is part of our concept of pain that certain patterns of behaviour (crying, moaning, sighing, gnashing one's teeth, holding or protecting the aching part of one's body, etc.) are natural expressions of pain (cf. *PI* §244). They are the typical forms of behaviour of someone in pain who is unrestrained and willing to show his feelings.

Wittgenstein here introduces a useful distinction between two types of evidence: *criteria* and *symptoms*. If it is part of the very meaning of a term '*F*' that some phenomenon is (good, though not infallible) evidence for the presence of *F*, then that phenomenon is a criterion of *F*. If, however, we have only discovered through experience that *F* is usually accompanied by a certain phenomenon, then that phenomenon is only a symptom of *F* (*BB* 24-5, cf. *PI* §354). The upshot of the ascribability argument (*PI* §302) is that in order to ascribe a sensation to a particular person we need a *criterion* by which to identify sensations in others (for it must be determined what it *means* for somebody to *have* a sensation). And there is indeed such a criterion, namely appropriate expressive behaviour, which is not just a symptom of sensations, but forms an integral part of the grammar of sensation words. But with that criterion in place the inner-object picture collapses. As there is a *conceptual* link between pain, for example, and certain forms of expressive behaviour, the idea that in spite of that behaviour — by reference to which the word 'pain' has been given its meaning (*PI* §244) — there might never be any pain becomes inconsistent.

It has to be admitted, of course, that the criterial link between the inner and the outer is not a neat one-to-one correlation. There is the possibility of deception: pain-behaviour may be insincere. However, the sceptical consequences of this qualification are limited, for four reasons:

First, deceit too is a distinctive form of behaviour (*NfL* 241d): there are criteria by which to tell whether someone is sincere or deceitful; otherwise we could not meaningfully speak of sincerity and deception (*LW* 42g). Although we may not always be able to tell whether someone is sincere or not, we can normally say what kind of further evidence would settle the matter.

Secondly, deceit is a rather complicated form of behaviour that can only be attributed to creatures whose behavioural repertoire displays a good deal of intelligence. It requires a motive based on an understanding of what is to be gained by simulation (*RPP I* §824). Thus the idea that a baby may be dishonest is incoherent (*PI* §249).

Thirdly, although for any isolated utterance or piece of behaviour one could imagine a context that would expose it as mere pretence, this is not true of sufficiently long sequences of circumstances and behaviour. If over months you observe someone suffering from an obvious and severe injury, it is ridiculous to insist that this might be a mere pretence.

Finally, that it is logically possible for a proposition to be false is not in itself a reason to doubt its truth. It is simply a trivial grammatical feature of *any* empirical proposition. (This is a general objection to all forms of philosophical scepticism.)

(iv) In one of the most discussed passages of the *Philosophical Investigations* Wittgenstein illustrates the way sensation language is envisaged by the proponent of the inner-object picture. To highlight the dualist belief that sensations are logically independent of expressive behaviour it is assumed in a thought experiment that I have no natural expression for a given sensation, but only have the sensation (*PI* §256). And now: 'I want to keep a diary about the recurrence of [that] sensation. To this end I associate it with the sign "S" and write this sign in a diary for every day on which I have the sensation' (*PI* §258).

Wittgenstein continues to object that when trying to keep such a private sensation diary (1) I have no criterion of correctness. Therefore, (2) whatever is going to seem right to me is right. And so, (3) here we can't talk about 'right' (*PI* §258). This is what I call the *no-criterion Argument*.

The final conclusion is a little exaggerated. 'Whatever is going to seem right to me is right' is not the same as, 'Anything is right' (which would make the use of the word 'right' pointless). It would still be *wrong* for me to write down 'S' on a day when none of my sensations seemed to be of the same kind as the one I initially called 'S'. What Wittgenstein should have concluded is: (3*) 'And that only means that here we can't talk about an *error*.' This is exactly what he says elsewhere about reporting one's pain (*PI* §288).

Now, what about conclusion (2)? Given that there is no criterion to check whether truthful 'S' inscriptions are *objectively* true or false (i.e. true or false independently of the person's impression), is it not possible nonetheless to wonder whether they are? After all, there are other cases of unverifiable propositions that must nevertheless be objectively true or false. For example, 'Immediately before her death Queen Victoria remembered her wedding day' is presumably a proposition that is either true or false, although as it happens it can never be confirmed or disconfirmed. The question arises whether the private diarist's 'S' inscriptions might not be of the same character: Could they not be objectively true or false, even though it is impossible to check?

There is, however, a crucial difference between those cases. In the case of the conjecture about Queen Victoria's last thoughts we may insist on truth or falsity because we know perfectly well what it means to say, for example, that someone remembers Prince Albert, etc. That is to say, although stymied in these special circumstances, *we are able in countless other cases to ascertain whether that predicate applies or not*. In contrast, with 'S' the problem is precisely that we cannot draw a distinction between a generally unproblematic predicate and its unverifiable application in only some particular cases. We cannot refer back to other applications of 'S' under more straightforward circumstances. The diarist's entries are, and will remain, the only applications of 'S' available. There is no room in this case for a notion of truth that would not coincide with sincerity, not simply because 'S' inscriptions are found to be uncheckable, but because (unlike in the case of Queen Victoria's last thoughts) there is no predicate involved whose applications would be checkable under *any* circumstances. Hence, the attempt to construe the sign 'S' as the name of objective mental occurrences fails.

It should be emphasised that this conclusion, (2), is not based on the controversial doctrine of verificationism: the view that the sense of a proposition is its method of

verification or falsification, and that therefore any un(dis)confirmable statement must be meaningless. For one thing, the issue is about simple *predicates*, not about statements. For another thing, the claim is not that because 'S' inscriptions are unverifiable they are meaningless; the claim is merely that 'S' inscriptions cannot be construed as descriptions of inner objects. After all, that 'S' inscriptions are not based on criteria and therefore unverifiable does not distinguish them from countless ordinary first-person present-tense psychological utterances. I cannot apply any criteria to verify my claim that I have a slight headache (nor can anybody else), and yet one would surely not want to dismiss that statement as meaningless.

The no-criterion argument provides yet another refutation of the inner-object model. If a sensation were an inner object perceived and identified through introspection, it should be possible occasionally to misperceive and misidentify it. One could, for instance, be mistaken in one's belief that one was in pain. That sounds absurd. For we do not identify our sensations by criteria (*PI* §290), and where there is no criterion, there is no possibility of error. But how *do* we identify our (own) sensations? The answer is that we do not identify them at all.

How then do words refer to sensations? Our natural responses to pain are not the result of an identification. Rather, the sensation of pain *makes* us cry. And later verbal expressions are grafted onto natural ones: we are trained until a suitable verbal expression comes as naturally to us as a cry or a moan (although, of course, we also learn to suppress all such reactions) (*PI* §244).

It is part of the very concept of pain that certain forms of behaviour are natural expressions of pain. If someone behaves in such a way, and is truthful — that is what we call 'being in pain'. That is the way the concept of pain is formed. The question whether someone is in pain or not is decided by that person's sincere behaviour. As a consequence, the philosophical query whether people's sincere avowals of pain are correct — whether they are *really* in pain when they truthfully says they are — does not make sense. Sincere avowals are correct 'by definition'; their correctness is built into our concept of pain.

In resolving the philosophical problem of other minds Wittgenstein draws attention to its connection with another one: the problem of first-person authority. As noted above, Wittgenstein's objection to Cartesian dualism is not (as one might expect) that it exaggerates the distinction between the mental and the physical, but, on the contrary, that it makes it appear too slight. The dualist conception of

psychological phenomena is too much shaped after our concept of physical objects or events. Construed as inner objects, sensations and thoughts in others become elusive; but, as Wittgenstein was the first to note, sensations and thoughts in ourselves become problematic too. If expressing one's feelings were a matter of observing and reporting inner occurrences, the special authority we have in expressing what we feel would be puzzling. Error would at least appear to be a possibility. Only when the inner-object model is rejected, the puzzle of first-person authority can be resolved. Our virtual infallibility (linguistic mistakes apart) in expressing our own feelings is not due to our eagle-eyed gaze of introspection, but simply to linguistic meaning. Whatever you find painful *is* painful to you. That is what we mean by 'painful'.[3]

4. Meaning and understanding: the paradox of the instantaneous experience of complex contents

Unlike bodily sensations (such as pain), understanding, thoughts, intentions and memories have a specific intentional content: they are *about* something else. How is that possible? The most natural, indeed virtually unavoidable, answer is that such mental occurrences must be inner *representations* of whatever their contents are. But this view, Wittgenstein shows, leads to a formidable problem: Understanding, intention, expectation, remembering, and other such mental occurrences, can have remarkably rich and complex contents. It may take very long to spell out completely what exactly someone understood, intended, expected or remembered on a given occasion. And yet, it appears that the understanding, intending, expecting or

[3] It is important to have a clear understanding of the target of the no-criterion argument. Contrary to what is often maintained in the secondary literature, the argument is not meant to be an objection to the private sensation diary described in §258. Rather, it is an objection to a misconstrual of this or indeed any other sensation language: namely the inner-object conception and its immediate consequence that one *perceives* one's own sensations. Hence, the no-criterion argument is also directed at the view that others can only have an indirect perception of my sensations: For if I do not *perceive* my sensations, others cannot be said to perceive them only indirectly. The point concerns ordinary sensations, such as pain, just as much as the 'private', fictitious scenario in which natural expressions or a familiar use of the sensation word are set to one side. What the argument is *not* supposed to show is that one cannot keep a record of sensations without a natural expression.

remembering can occur instantaneously: in a flash. How is that possible? How can so much be experienced in a flash?

Consider, for instance, that it is possible to grasp the meaning of a word in an instance (*PI* §§138-9), although an account of the meaning of a word, of all the details of its use, would have to be extremely long and complicated. How can it be present to my mind all at once? 'What really comes before our mind, when we understand a word?', Wittgenstein asks, and proceeds to consider the Lockean view that it is 'something like a picture' (*PI* §139). However, understanding a word cannot be the same as having a mental image, for any image can be interpreted, or applied, differently. For instance, the perspectival drawing of a cube (as a candidate for giving meaning to the word 'cube') may also be taken as a two-dimensional figure consisting of a square and two trapeziums. Again, the image of a dog may be taken to represent: a particular golden retriever, or any golden retriever, or any dog, or a mammal, or an upright position, or many other things (cf. *PI* §139). Indeed, one cannot even assume that when an image is produced in one's mind by hearing a word this image must represent what one takes the word to mean. The word 'winter' may produce in me the image of an old aunt pouring out tea without my being under the strange misapprehension that that is what the word 'winter' means. Two people hearing a word can have the same mental image and yet a different understanding of the word, manifested in the different ways they apply it and the different explanations they give (cf. *PI* §140). And it is also possible that two people have different mental images when hearing the same word although they are in perfect agreement about its meaning. These objections are equally applicable to the complementary account of meaning something as having mental images (*PI* §663).

So, we may indeed have a mental image of a cube in our mind when hearing (or uttering) the word 'cube', but that is not what understanding (or meaning) the word consists in, as one can understand without having any mental image and no mental image guarantees understanding. Still, the idea is there, and persistently recurs in the *Investigations*, that a mental state, possibly instantaneous, can have a determinate content, possibly very extensive and complicated. After all, we do mean, understand, remember or intend complicated processes, for example:

> There is no doubt that I now want to play chess, but chess is the game it is in virtue of all its rules (and so on). Don't I know, then, which game I wanted to play until I *have* played it? Or is it rather that all the rules are contained in my act of intending? [*PI* §197]

This is what I call the *paradox of the instantaneous experience of complex contents*: The contents must all be there in a flash, for I can correctly avow that at a particular moment I intend (or have understood, or expect, or remember) something. But then again, the contents are not all there in a flash, for I am not really aware of all the details: all possible uses of the word; or all rules of chess; they are not all in front of my mind at the same time.

Consider the example of a pupil taught to write down (the beginning of) the series of even numbers on being given the order '+ 2'. Up to 1000 he does that correctly; we ask him to continue and 'he writes 1000, 1004, 1008, 1012' —:

> We say to him: "Look what you're doing!" — He doesn't understand. We say: "You should have added *two*: look how you began the series!" — He answers: "Yes, isn't it right? I thought that was how I *had* to do it." — Or suppose he pointed to the series and said: "But I did go on in the same way." — It would now be no use to say: "But can't you see . . . .?" — and go over the old explanations and examples for him again. [*PI* §185]

Obviously, the pupil did not understand the order '+ 2' the way his teacher understood it and meant it. The pupil was not meant to write 1004 after 1000. Does that mean that the teacher had thought explicitly of 1002 as the correct number to write after 1000? No, or if he had thought of that particular transition, then not of countless others. He rather thought that the pupil 'should write the next but one number after *every* number that he wrote' (*PI* §186). This is a general rule, or formula, that may have been in the teacher's mind when he gave the shorthand order '+ 2'. Now the question is whether the presence of such a rule in someone's mind can account for his understanding or meaning an infinite series. Obviously not (cf. *PI* §152). The teacher may give the rule to his pupil (instead of the abbreviated signal '+ 2'), make him learn it by heart, and yet there is no guarantee that the pupil will continue correctly. It is conceivable that he misunderstands the explicit rule just as he misunderstood the

order '+ 2': taking it as *we* would take the rule: 'Add 2 up to 1000, 4 up to 2000, 6 up to 3000, and so on' (*PI* §185). So, any rule, even the most explicit one, can be misunderstood; and in endless ways, too: whichever way the pupil continues the series, his writing can always be regarded as in accordance with the rule — *on a suitable interpretation* (cf. *PI* §86). And now it is puzzling how it should be at all possible to follow a rule, for 'Whatever I do is, on some interpretation, in accord with the rule' (*PI* §198)!

This line of argument (which Wittgenstein does not endorse, but present for critical discussion) can be put as follows:

1) A series of numbers can be continued in various ways.

2) The way it is meant to be continued is expressed by a rule (or formula).

3) A rule (or formula) can be interpreted in various ways.

4) Therefore, a rule (or formula) cannot determine a continuation.

5) Therefore, our meaning (as expressed by a rule) cannot determine a continuation.

The final conclusion (5) is partly based on (2): That the teacher *meant* the series to be continued in a particular way we tried to explain by saying that he had some general rule or formula in mind. So, if that rule cannot fix a standard of correctness, clearly, his meaning that rule cannot do it either.

The flaw (to which Wittgenstein wants to direct out attention) lies in the step from (3) to (4): That a rule can be interpreted in various ways does not entail that a rule cannot determine a continuation. One might as well argue as follows:

Your bicycle could always be stolen.

Therefore, it can never be used.

It is obviously fallacious to argue from 'It can go wrong' to 'It can't go right'. *If* your bicycle is indeed stolen from you, then of course you cannot use it; but it is well possible, perhaps even likely, that it is not stolen and then you *can* use it. Moreover, even if your bicycle is stolen, it can (and probably will) still be used — though not by you. Similarly: for any given continuation (e.g. 1000, 1002, 1004), a suitable rule (e.g. $x_n = 2n$) can always be interpreted *not* to yield that result — but it need not be so interpreted and normally it isn't. And even if the rule '$x_n = 2n$' is interpreted in a deviant way, it will still determine a continuation of the series, albeit not the one we expected (but perhaps the one written down by the deviant pupil).

There remains the worrying consideration that a rule needs to be interpreted at all. It cannot *by itself* determine a continuation. It seems that in order to understand it in some way, you need to give it an interpretation. And once you accept that an interpretation is necessary, you seem to be launched on an infinite regress: for whatever interpretation you give, it is in no better a position than the rule itself. It, too, needs to be interpreted in a particular way (cf. *PI* §86) — and so on, and so forth.

Wittgenstein's response to that puzzling consideration is that for us to think, even *for a moment*, that on a certain understanding the formula '$y = x^2$', for example, yields 25 for x = 5, the infinite regress that seemed to threaten must have been stopped. Our understanding in this case cannot just be an interpretation: that is, another formula to paraphrase the first, of which again it would be an open question how to understand it (*PI* §201). Our understanding of a rule cannot forever be mediated by another rule.

What it amounts to is simply this: we are able to work out that for x = 5 the formula '$y = x^2$' yields 25. How? Well, 'I have been trained to react to this sign in a particular way, and now I do so react to it' (*PI* §198). This is the core of the matter: knowing-how (a skill) cannot, ultimately, be explained in terms of knowing-that (a piece of information). For any piece of information, for any formula in the mind, we would again need to *know how* to apply it. It may happen occasionally that our understanding of a formula is based on, or mediated by, our understanding of another formula. But such a translation of one formula into another cannot be the basic case of understanding: or else we would never *begin* to learn the meaning of any formula. Likewise, our basic linguistic understanding, the mastery of our first language, can evidently not be explained in terms of translation into another language.

There is a strong temptation to think that when we have understood in general how to proceed, somehow all the correct applications must be laid down in advance in the mind, so that each step can be justified by reference to that perfect mental instruction manual. But that perfect mental instruction manual is an illusion and (as explained) a logical impossibility. Our basic skills must stand on their own. They are engendered by training, but what they enable us to do is open-ended and cannot be exhaustively listed in training, nor could there be such an exhaustive list in the mind or anywhere else (cf. *Z* §§300-1).

The paradox of the instantaneous experience of complex, or even infinite, contents resulted from a mistaken, but very natural, idea of a mental process or

occurrence. It is extremely tempting to envisage mental occurrences as comprehensive representations: that must somehow *contain* everything the mental representation is about or directed at. That kind of mental representation would be astonishing enough where what we have in mind is fairly complex — as with the rules of chess, which appear to be represented in our momentary desire to play chess (*PI* §197); but it becomes patently impossible when we mean or understand an infinite arithmetical series. At first glance it seemed that such infinity could conveniently be represented by a short formula, but then it became clear that our meaning or understanding the formula would have to contain how the formula was to be applied in an infinity of instances (to forestall an infinity of possible misunderstandings) — which again, appears to make such meaning (or understanding) a mind-boggling feat. So, how can the mind perform such a feat? It cannot. The truth is that no such marvellously rich representation occurs; and it is an error to think that it needs to occur for meaning (or understanding) to be possible. That is Wittgenstein's dissolution of the paradox of the immediate experience of complex, or even infinite, contents.

But then: if meaning one thing rather than another is not brought about by a comprehensive and infallible mental representation of that thing, how *is* it brought about? — In response to this question Wittgenstein merely invites us to remember under what circumstances it is correct to say that somebody meant something. There is no such thing as an intricate mental mechanism of meaning that could be investigated and explained (*PI* §689). There is no process of meaning: 'For no *process* could have the consequences of meaning something' (*PI*, p.218: PPF §291). Any process you might think of can occur without your meaning the thing in question. What makes it true then that when I say 'Let's play a game of chess' I mean the game of chess with all its rules, and not some other game? The link is not made by some miraculous mental mechanism, but by the circumstances in which the utterance is made: In English 'chess' is the name of a particular game the rules of which have been listed and widely published; there are clubs and organisations that insist on those rules, and people are taught those rules at home or at school. So when a competent and normally educated English speaker speaks of 'chess' he can be taken to mean what we all call 'chess' (cf. *PI* §197). In other words, there is no doubt that when he speaks of 'chess', he speaks of chess (cf. *PI* §687). — Admittedly, it is conceivable that someone uses the word 'chess', but means backgammon; either because his

knowledge of English is less than perfect, or simply through a slip of the tongue. But how would such a person differ from someone who uses the word 'chess' correctly? There need not be any difference in what went through their heads at the time of the utterance: If God had looked into their minds he might not have been able to see *there* which game they wanted to play (*PI*, p.217: PPF §284). Rather, to say that they meant different games amounts to saying that they will respond differently when, say, the pieces are set up in front of them, or that they *would* have responded differently, had the matter been pursued further (*PI* §§187, 684). Frequently, the difference between different thoughts or intentions is only a conditional one: one that *would* manifest itself (or would have manifested itself) under certain circumstances.

5. Thinking

What is thinking? Again, the most natural answer will be informed by the inner-object model. Thinking appears to take place in our minds, and so in order to get clear about its nature, or about the meaning of the word 'thinking', we watch ourselves, the contents of our minds, while we think, expecting that 'what we observe will be what the word means'. Wittgenstein objects that 'this concept is not used like that' (*PI* §316). What can be observed in my consciousness when, for example, I am thinking that I have to write to my aunt thanking her for a birthday present I didn't like at all? Not a lot. If it is not the first time that I think of this mildly tiresome obligation, the thought will perhaps re-appear in my mind as nothing more than a peculiar feeling of uneasiness. I don't have to repeat to myself what I am uneasy about; I know straightaway what this feeling is about. When I dwell a little longer on the thought, I will perhaps see a mental image of the parcel I received from her together with a vague auditory image of the telephone conversation we had six months ago. Are these images the thought? No, for if I just told someone the images, he would not get the thought (*Z* §239). Even if I add, what is not really part of the mental images, from whom I had the parcel and with whom I had spoken on the telephone, it is impossible to work out that I am thinking of writing her a thank-you letter. Exactly the same mental images might accompany somebody's thought that he

would *not* write in response to the parcel, or that he expected soon to receive another parcel from the same sender.

We are inclined to believe that when we think something the content of the thought must somehow be represented in our consciousness, but this is usually not the case. And even when the contents of a thought are fairly comprehensively expressed *in foro interno*, one cannot simply read off the thought from this inner representation. For the fact that the *words* 'I have to write to my aunt to thank her for that awful present' go through my mind, does not necessarily mean that I am having that thought. Just now, for example, they went through my mind when I did not have any such thought — merely as an example for this philosophical consideration.

The word 'think' functions rather differently from the word 'write'. When I am *writing* that it will rain tomorrow, this content is completely laid down on the paper and can be read off there; when I am *thinking* that it will rain tomorrow there need not be any such readable representation in my mind. Usually, what can be found there resembles more a private shorthand, for example, a few diagonal lines which *I* take *in this case* as indicating that it will rain tomorrow, although that is not a meaning the lines have for others or in other cases.

> I can see, or understand, a thought complete before my mind's eye in a flash in the same sense in which I can make a note of it in a few words or a few pencilled dashes. [*PI* §319]

That of course raises the question: 'What makes this note into an epitome of this thought?' (*PI* §319). The answer is: 'The use that I make of it' (MS 124, 218). Should anyone ask me what those pencilled dashes mean, I can explain it; and if I look at the note later it will remind me that I expected rain for the next day. Of course, my mental images cannot be presented to others nor kept for my own later recollection. But the analogy lies in the fact that I can say authoritatively and without being constrained by semantic conventions what the images or words flitting through my mind mean: what thoughts they are illustrating for me. This is the way our concept of thinking functions: What I think is what I can sincerely declare to be my thought, and what can also manifest itself in my further behaviour. The words and images that go through my mind when I am thinking are ultimately quite irrelevant.

In his early philosophy, in the *Tractatus Logico-Philosophicus*, Wittgenstein had taken thinking to be a process accompanying speech and giving it meaning, a process that was itself rather similar to speaking. In the *Investigations* he was to criticise that view:

> When I think in language, there aren't 'meanings' going through my mind in addition to the verbal expressions: language itself is the vehicle of thought. [*PI* §329]

The idea of thinking as a process that accompanies speech and gives it significance doesn't sit well with the fact that it is possible to think aloud. When I think aloud: 'What do I have to do today? Ah, yes, write to Aunt Agatha to thank her for that awful present she has sent me.' — what is it that gives meaning to those words? Are the words accompanied by a process of thinking or meaning them? In that case, the actual thinking would be the accompanying process, not the words. And if one always needed the backing of such a mental process, it wouldn't be possible at all to think aloud. But then it wouldn't be possible to think quietly in words either. For surely, whether the words are uttered aloud or are only articulated quietly in my mind cannot make any difference. But in fact, when I said those words there was nothing else going on in my mind at the time. It *is* possible to think in words, aloud or quietly, without any accompanying mental process.

Having ascertained that thinking can take the form of words uttered aloud, Wittgenstein raises the question whether one can imagine people who could *only* think aloud (*PI* §331). It is not difficult to imagine that everything somebody says to himself he says aloud, like a person on stage whose thinking is thus laid open to the audience. In that case he could not simultaneously speak to others and articulate thoughts to himself. But that is perhaps something most people cannot do anyway. I find my own mental comments are always made between public utterances and not simultaneously with them. So the transition from speaking quietly in one's mind to speaking aloud seems unproblematic. The crucial question, however, is whether it is possible for *all* thinking to be articulated in words.

In that case our intellectual life would have to slow down considerably. For in fact, our thoughts often occur much faster than we could express them in words. Somebody who could only think in words would take much longer to have a thought

in its entirety before he could act accordingly. There are many situations, for example when driving a car, where such a delay in one's thoughts and actions could have serious consequences and be a great handicap.

However, the philosophical concern goes deeper than that. A bit further in the discussion, Wittgenstein raises the following point:

> But didn't I already intend the whole construction of the sentence (for example) at its beginning? So surely it already existed in my mind before I uttered it out loud! [*PI* §337]

However much a person is inclined to think aloud or soliloquise, it is hardly possible that when uttering a word of his thoughts he never knows what words will come afterwards. If one were to speak the beginning of a sentence without any anticipation of the following words one could not know why one uttered the first words, which frequently don't make any sense without the sequel. In that case one would have to speak as in trance. In fact, however, somebody who consciously expresses his thoughts (or thinks aloud) does not need to wait for the end of his sentences to find out himself what he was going to say. In some sense, the thoughts are already in one's mind before they are fully articulated. It follows that, even when one is thinking in words, the thoughts cannot really be identified with their verbal expression. My awareness of my thoughts is not read off, and not entirely simultaneous with, the words that may embody the thoughts.

This is closely connected with the considerations of the previous section, which were triggered by the puzzlement at the fact that I can in a split second have the intention to play chess, for example. How can something as complicated as the game of chess be all in my mind in such a short space of time? The same problem arises when I take the time to formulate the sentence 'I would like to play chess'. For I have to utter these words *with understanding*: that is, I must while I am speaking know what I mean by 'chess'. All the numerous details of what I mean (e.g., that it is a game in which every player has eight pawns that can move only forward) are part of my thought. (For I didn't mean, for example, a game in which every player has nine pawns.) In other words, even when I am thinking in complete sentences, my thoughts comprise more than the mere words. They also comprise a certain understanding of those words: the capacity to explain them and to develop their countless implications.

This is the reason why thinking cannot strictly be identified with any process in the mind: thinking requires an understanding of its contents, a capacity to explain it, apply it, and draw inferences from it. Yet a capacity belongs to a different logical category from a process.

6. Voluntary action

What is a voluntary action? How is it to be distinguished from a mere event? To begin with, a voluntary action is performed by a conscious agent. However, that is not sufficient to characterize voluntariness, for conscious agents can also be passively involved in a mere event. I can jump off a wall, but I can also inadvertently fall off it or be pushed off it against my will. The bodily movements in these cases *need* not be different. Suppose in a given pair of cases they are exactly the same, how then are we to distinguish between the voluntary action and the mere event? It seems that the voluntary action must contain *more* than the mere event: there must be an extra element of willing or intention. So it should be possible to isolate that element of willing by a thought experiment of subtraction: 'what is left over if I subtract the fact that my arm rises from the fact that I raise my arm?' (*PI* §621). What is Wittgenstein's answer to this question? — He does not give an answer. As so often in philosophy, the question is misguided, and instead of answering it we should find out what is wrong with it. It is, in fact, another instance of referentialism inviting us to construct some spurious inner object (event, process, or state). Words used to characterize an action as voluntary (like 'will' and its cognates) are uncritically taken to denote some mental occurrence, some phenomenon that added to the mere bodily movement turns it into a voluntary action.

This is the inner-object model of voluntary action. According to its classical version, going back to Descartes and the British Empiricists, 'what is left over if I subtract the fact that my arm rises from the fact that I raise my arm' is an *act of will*, or (as Hume called it) a *volition*. For a bodily movement to be voluntary it must be *caused* by an act of will; without such a cause the same movement would be involuntary. Wittgenstein was familiar with similar views held by Bertrand Russell and William James. He offered three objections to this theory:

The *first objection* is very simple: The volitions or acts of will postulated by the theory do not exist. If we take an impartial look at what goes on in our minds whenever we move our body voluntarily, no suitable mental events causing the movements come to light. However, the elusiveness of acts of will tends to be obscured by philosophers' selective attention, when they focus on only a few especially favourable examples, such as this one: 'I deliberate whether to lift a certain heavyish weight, decide to do it, I then apply my force to it and lift it' (*BB* 150). Here we have some occurrences that could, without absurdity, be thought to constitute willing: some anticipatory thinking of the action, an act of resolve, a sensation of bodily effort. And now we take our ideas about voluntary action from this kind of example and assume lightly that those ideas must apply to all cases of willing (*BB* 150). But of course not all cases are like that. We frequently do things without any such preliminaries. Just think of ordinary speech, which is often entirely unpremeditated and effortless, yet not for that matter involuntary.

Second Objection: Willing is thought to be a mental occurrence, but a mental occurrence must be either voluntary or involuntary. That leads to a fatal dilemma: If the mental act of willing is itself subject to the will, in order to be proper willing it would have to be willed. But then we are launched on an infinite regress: For the event of willing to be voluntary it has to be caused by an earlier event of willing; but that earlier event, too, in order to be voluntary would have to be caused by yet an earlier event of willing, and so on *ad infinitum* — which is absurd. So it seems more promising to deny that willing itself could be subject to the will: 'I can't will willing' (*PI* §613). But that sounds odd as well. It would appear that 'willing too is merely an experience … It comes when it comes, and I cannot bring it about' (*PI* §611). But now the whole idea of voluntariness, of being in control of one's actions, seems to be lost. That must be wrong too (*PI* §612). The dilemma shows that the whole question (whether or not willing can be willed) is misbegotten. Willing is not the sort of thing of which it makes sense to ask whether it is voluntary or involuntary. 'Willing' is neither the name of an action, nor of a passive experience. It is not the name of a mental occurrence of any kind.

Third Objection: According to the inner-object model, a voluntary bodily action is a bodily movement caused by a mental act. Thus, on this theory I *bring it about* that, say, my arm rises. But in fact, Wittgenstein objects, I don't (*PI* §614). I don't do anything else as a means to effect the rising of my arm. In particular, it

cannot be said that I contract certain muscles in order for my arm to go up, for I don't even know which muscles need to be contracted for the arm to go up. (It is rather the other way round: I could raise my arm in order to bring about the contraction of whatever muscles are involved in the process.) Nor do I bring about bodily movements by acts of wishing or deciding. Wishing that something may happen is actually incompatible with doing it voluntarily (*PI* §616). The word 'wish', like 'hope', implies that one is not fully in control of what will happen. If I wish my arm to rise and, lo! it does — it wouldn't be my own action and I'd be very surprised (*Z* §586b). A decision to raise my arm, on the other hand, is of course likely to lead to my raising my arm; but it does not just cause my arm to go up. Again, I'd be rather surprised if it did. It would not be my own doing (*PI* §627). A decision to do something may of course lead to a voluntary action, but it occurs *before* the action and cannot be regarded as part of it. Hence it cannot figure in the analysis of the concept of a voluntary action.

Those three objections show that the inner-object model of voluntary action must be rejected. Words like 'voluntary' or 'willing' do not stand for some distinctive mental occurrence that must precede or accompany a movement for it to be voluntary. How, then, is the word 'voluntary' used? Again, we should not expect the answer to be an exciting revelation. The concept is a familiar one, so its philosophical elucidation can only be a reminder of what in practice we are all familiar with. 'Voluntary movement is marked by the absence of surprise' (*PI* §628). I am not a third-person observer to my own behaviour: I cannot look on with interest to see what will happen next, and then perhaps be surprised by it. That is related to the observation that my action's being voluntary is incompatible with my wishing for it to happen (*PI* §616). For one can only have wishes about what is not entirely under one's control, and where something is not under one's control one can doubt whether it will happen (or never have thought of it), and hence be surprised if it does. Of course, that is not to say that all things that happen to us come as a surprise; but with mere events and involuntary actions surprise is at least always logically *possible*, whereas to the extent to which an action is voluntary there is logically no room for surprise.

Why not? It is tempting to think that one is not surprised here because one knows so reliably of one's own voluntary movements. Then naturally the next question is: *how* does one know, and the almost unavoidable answer is that one *feels*

one's own voluntary movements, perhaps in one's muscles and joints: '"How do you know that you have raised your arm?" — "I feel it."' (*PI* §625). But feelings can be deceptive. Whatever sensations may be characteristic of raising one's arm, it is surely conceivable that in a laboratory they might be produced artificially, by drugs or electric currents. So when now I raise my arm with my eyes shut, whatever sensations I have in my muscles and joints it should be conceivable that *they* are deceptive. Hence, if my awareness of my voluntary movements were based on such sensations, I should be able in this situation to consider it possible that I am *not* moving my arm (*PI* §624). But I find myself unable to do so; for my certainty that I am moving my arm is not based on the evidence of such sensations. I am just certain that I have raised my arm, and there is no evidence on which my certainty is based (*PI* §625).

The puzzle of first-person authority about one's own agency is not unlike that about one's sensations, and Wittgenstein dissolves it in a similar way. The puzzle is generated by treating the case as one of *knowledge*; which, first, makes it appear strange that there shouldn't be any possibility of error, and which, secondly, makes us look (in vain) for some grounds or evidence on which this extraordinary knowledge could be based. To remove the puzzle we only need to realize that the certainty is the result of our grammar. It is built into our very concept of a voluntary action that the agent is aware of it (cf. *Z* §600) — which is therefore as unsurprising as the fact that bachelors are without exception unmarried.

Consider a related case: You express an intention to go for a walk. Now the question 'How do you know?' makes no sense (or could only be understood as asking: 'How do you know that you will not be prevented?'). An expression of intention is not based on any evidence and cannot be erroneous. In this respect expressions of intention are like declarations that one acted voluntarily: the agent's authority is simply built into our concepts.

7. Wittgenstein's influence

Wittgenstein's anti-Cartesian insistence that there is a conceptual link between mental states and expressive behaviour was often accused of being a form of behaviourism. But in fact, he nowhere showed any inclination to try to *reduce* mental states to

behaviour (or behavioural dispositions), and it is indeed fairly obvious that for most mental phenomena there is no immediate expressive behaviour. Thoughts, for example, manifest themselves in words or deeds only when suitably combined with other mental states or dispositions. Taking this interrelatedness of mental states seriously, while accepting Wittgenstein's valuable insight in the importance of behaviour for the explanation of the mind, led philosophers in the 1970s from behaviourism to a more sophisticated theory: functionalism.

How do Wittgenstein's views on the mind compare and relate to functionalism? Certain similarities are undeniable. When Wittgenstein in order to clarify our concept of pain sketches the case of a child that has hurt himself and cries, and is later taught verbal pain-behaviour (*PI* §244), one may be tempted to see that as an anticipation of the functionalist conception of a human being as a mechanism that connects certain inputs (injury) with certain outputs (crying). Teaching may change this input-output function, presumably by affecting the mechanism's inner states that mediate between input and output.

As explained in the first section, for Wittgenstein, the only licit philosophical method was conceptual analysis. Yet conceptual analysis is not likely to yield a simple and yet non-trivial reductive formula about the nature mental states. In particular, conceptual analysis provides no reason to expect that our concepts correspond to the functional states of a determinate input-output mechanism. For example, given my strong desire to eat strawberries, it may be very *likely* that coming to know that there are strawberries for me in the fridge will cause me to go and eat them, but then again, one can easily imagine that I won't:— I may decide to keep them for later; or to offer them to my neighbour; or I may suddenly remember that I need to make an urgent phone call; or I refuse them because I don't want to be under any obligation to the giver; or it may occur to me that I don't really care for strawberries any more. And even if there is no understandable reason for me not to eat the strawberries, it remains certainly conceivable that I don't. Sometimes people do react in strange ways that they can't fully explain. Although there are conceptual links between our descriptions of beliefs, desires and actions, they don't normally allow any predictions with logical certainty. All we can say is that a certain desire and a certain belief make a certain action *understandable*; or: under the circumstances a certain belief was a good *reason* to act in a certain way. But other actions would

have been equally understandable and justifiable by reasons; and even actions that are not readily or fully understandable would at least be conceivable.

Note also that, with the possible exception of sense perception, our concepts of mental states are virtually open with respect to the causal origin of those states. If my neighbour's ironic smile caused in me the firm conviction that he was a Russian mafia boss, I might well be accused of being irrational; yet this insufficient ground for my belief would not speak against attributing this belief to me if I expressed it with all seriousness. It is not built into our concept of a belief that it has to be caused in certain specific ways. Again, we have of course some empirical knowledge about the likely causes of pain and might be sceptical if somebody complained of pain in the absence of any such likely cause; yet its not a *conceptual* truth that for something to qualify as pain it must be caused in such and such a way. Indeed, we know from experience that occasionally people suffer pain from unidentifiable psychological causes.

So a *specific* functionalist identity claim can certainly not be established *a priori*. Our concepts of mental states do not imply any specific possibilities of origin or causal potential. Is it then perhaps *a priori* plausible to assume that any mental state is identical with *some* specific functional state defined by its possible origin and causal potential, to be discovered by future psychological research? No, for it cannot be ruled out that two instances of the same mental state may differ in their possible origin and causal (or dispositional) potential. Thus, some instances of the belief that *p* may in a certain surrounding of other mental states cause a feeling *s*, while other instances of the same belief do not. In fact, since we don't know the specific causal (or dispositional) potential of any mental states it is *a priori* unlikely that our concepts should classify mental states exactly according to their causal (or dispositional) potential. It appears even more unlikely if we bear in mind that it is an essential feature of our most common mental or psychological concepts that they are self-ascribable in a way that is both authoritative and not based on evidence. My sincere avowal of a feeling, a preference or a belief is, *ipso facto*, an expression *of what I feel, prefer or believe*, even though it is not based on any observation of my mind, let alone a study of its causal mechanisms. My sincere avowal would, obviously, itself be an effect of the mental state in question; an effect that, under normal circumstances, suffices for us to identify that mental state. So, our concepts of such *mental states* are such that we identify them by *one* telling effect, namely the subject's avowal. But

isn't it highly improbable that a *functional state*, defined by a list of *all* its possible causes and possible effects, should be identifiable from only a single one of its effects? Clearly, one should expect that this single effect, the subject's sincere avowal, could be caused by very many different functional states. The concept of a functional state is far more finely grained and sharply defined than that of, say, the belief that Paris is the capital of France, or that of an admiration for Daniel Auteuil.

Functionalism starts out with certain conceptual truths, namely that mental states are affected by perception and sensation; they affect or condition other mental states; and they lead to, and manifest themselves in, certain forms of behaviour. As far as these truisms (and their importance for the philosophy of mind) are concerned, there is agreement between Wittgenstein and functionalism. But then, functionalists are greatly impressed by the fact that these truisms highlight a certain similarity between a human being and a computer, a so-called Turing machine, with input, distinct functional states, and output. And now functionalism presses this analogy: insisting that it is not merely an analogy, but an identity. That, in fact, a human being *is* a Turing machine, albeit a highly complicated one, and that mental states just *are* functional states. This second step — from an analogy to an identity claim — is where functionalism parts company with Wittgenstein. This second step is neither the result of conceptual analysis, nor supported by empirical evidence. It is a typical philosophers' mistake. One is enthralled by a neat and attractive *picture*.

Here, as in other cases, Wittgensteinian considerations have been taken up by later philosophers of mind, but in a very different spirit, which quickly brings them in conflict with Wittgenstein's approach. His conception of philosophy limiting itself to conceptual clarification has not found many followers. Too strong seems to be the allure of quasi scientific theory construction in philosophy.[4]

8. Further Reading

[4] I am grateful to Amy Kind and an anonymous referee for their helpful comments on an earlier version of this paper.

M. Budd, *Wittgenstein's Philosophy of Psychology*, London: Routledge, 1989; new ed. 2014.

P.M.S. Hacker, *Wittgenstein: Meaning and Mind. Vol. 3 of an Analytical Commentary on the Philosophical Investigations*, Oxford: Blackwell, 1990.

P.M.S. Hacker, *Wittgenstein's Place in Twentieth-Century Analytic Philosophy*, Oxford: Blackwell, 1996.

P.M.S. Hacker, 'The Development of Wittgenstein's Philosophy of Psychology', in J. Cottingham & P.M.S. Hacker (eds), *Mind, Method, and Morality: Essays in Honour of Anthony Kenny*, Oxford: OUP, 2010; 275-305.

N. Malcolm, 'Wittgenstein on the Nature of Mind', *American Philosophical Quarterly*, vol. 4 (1970), 9-29.

T.P. Racine & K.L. Slaney (eds), *A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology*, Basingstoke: Palgrave Macmillan, 2013.

B. Rundle, *Mind in Action*, Oxford: Clarendon Press, 1997.

S. Schroeder (ed.), *Wittgenstein and Contemporary Philosophy of Mind*, Basingstoke: Palgrave Macmillan, 2001.

S. Schroeder, *Wittgenstein: The Way Out of the Fly-Bottle*, Cambridge: Polity, 2006.

9. Bibliography

David Hume (1748), *Enquiry Concerning Human Understanding*, eds: L.A. Selby-Bigge & P.H. Nidditch, Oxford: OUP, 1975.

Severin Schroeder (2006), *Wittgenstein: The Way Out of the Fly-Bottle*, Cambridge: Polity.

Ludwig Wittgenstein:

BB *The Blue and Brown Books*, Oxford: Blackwell, 1958.

BT *The Big Typescript, TS 213*. Edited and translated by C. Grant Luckhardt and Maximilian A. E. Aue. Oxford: Blackwell. (BT)

LW *Last Writings on the Philosophy of Psychology*, eds.: G.H. von Wright, H. Nyman; tr.: C.V. Luckhardt, M.A.E. Aue, Oxford: Blackwell, 1982.

MS Manuscript (numbered in accordance with G.H. von Wright's catalogue, in his 'The Wittgenstein Papers', in his: *Wittgenstein*, Oxford: Blackwell, 1982) from *Wittgenstein's Nachlass: The Bergen Electronic Edition*. Edited by the Wittgenstein Archives at the University of Bergen. Oxford: Oxford University Press, 2000.

NfL *Notes for Lectures on "Private Experience" and "Sense Data"* (1934-36), in PO, 202-88.

PG *Philosophical Grammar*, ed.: R. Rhees, tr.: A.J.P. Kenny, Oxford: Blackwell, 1974.

PI *Philosophical Investigations*, 4th ed.: P.M.S. Hacker & J. Schulte; transl.: G.E.M. Anscombe, P.M.S. Hacker, J. Schulte, Oxford: Wiley-Blackwell: 2009.

PO *Philosophical Occasions 1912-1951*, eds: J. Klagge & A. Nordmann, Indianapolis: Hackett, 1993.

RFM *Remarks on the Foundations of Mathematics*, eds: G.H. von Wright, R. Rhees, G.E.M. Anscombe; tr.: G.E.M. Anscombe, rev. ed., Oxford: Blackwell, 1978.

RPP *Remarks on the Philosophy of Psychology*, 2 vols, eds.: G.E.M Anscombe, G.H. von Wright, H. Nyman; tr.: G.E.M Anscombe, C.V. Luckhardt, M.A.E. Aue, Oxford: Blackwell, 1980.

TLP *Tractatus Logico-Philosophicus*, translated by D. F. Pears and B. F. McGuinness. London: Routledge & Kegan Paul, 1961.

Z *Zettel*, eds: G.E.M. Anscombe & G.H. von Wright, tr.: G.E.M. Anscombe, Oxford: Blackwell, 1967.