Research Repository UCD

Title	Artificial Intelligence and Wittgenstein
Authors(s)	Casey, Gerard
Publication date	1988-06
Publication information	Casey, Gerard. "Artificial Intelligence and Wittgenstein." Philosophical Society at St. Patrick's College, June 1988. https://doi.org/10.5840/philstudies19883239.
Publisher	Philosophical Society at St. Patrick's College
Item record/more information	http://hdl.handle.net/10197/5532
Publisher's version (DOI)	10.5840/philstudies19883239

Downloaded 2026-05-02 00:29:39

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

© Some rights reserved. For more information

Artificial Intelligence and Wittgenstein[1]

Gerard Casey
School of Philosophy
University College Dublin
gerard.casey@ucd.ie
www.ucd.ie/philosophy/staff/casey_gerard.htm

1. Introduction

The association of Wittgenstein's name with the notion of artificial intelligence is bound to cause some surprise both to Wittgensteinians and to people interested in artificial intelligence. After all, Wittgenstein died in 1951 and the term artificial intelligence didn't come into use until 1956 so that it seems unlikely that one could have anything to do with the other. However, establishing a connection between Wittgenstein and artificial intelligence is not as insuperable a problem as it might appear at first glance. While it is true that artificial intelligence as a quasi-distinct discipline is of recent vintage, some of its concerns, especially those of a philosophical nature, have been around for quite some time. At the birth of modern philosophy we find Descartes wondering whether it would be possible to create a machine that would be phenomenologically indistinguishable from

---

[1] This paper was originally delivered at the 1989 Spring Meeting of the Irish Philosophical Society. I am grateful to members of the Society for their comments and suggestions, some of which I have incorporated into the paper. To prevent a proliferation of footnotes, I am placing the bulk of the references to Wittgenstein's works within the body of the paper, using the following abbreviations. [See V. A. and S. G. Shanker (eds), *Ludwig Wittgenstein: Critical Assessments*, Vol. 5 (London: Croom Helm, 1986)].
PG: *Philosophical Grammar* (ed. Rush Rhees, trans. Anthony Kenny), Oxford: Basil Blackwell, 1974.
BB: *Preliminary Studies for the 'Philosophical Investigations' generally known as 'The Blue and Brown Books'*, New York: Harper & Brothers 1958—
RFM: *Remarks on the Foundations of Mathematics*, (eds. G. H. von Wright, R. Rhees, G. E. M. Anscombe, trans. G. E. M. Anscombe), Oxford: Basil Blackwell 1956.
PI: *Philosophical Investigations* (eds. G. E. M. Anscombe and Rush Rhees, trans. G. E. M. Anscombe), 2nd edition Oxford: Basil Blackwell, 1958.
RPP I: *Remarks on the Philosophy of Psychology, Volume I*, (eds. G. E. M Anscombe and G. H. von Wright, trans. G. E. M. Anscombe), Oxford- Basil Blackwell, 1980.
RPP II: *Remarks on the Philosophy of Psychology, Volume II*, (eds. G. H. von Wright and H. Nyman, trans. C. G. Luckhardt and M. A. E. Aue) Oxford- Basil Blackwell, 1980.
LW: *Last Writings on the Philosophy of Psychology*, Volume I: Preliminary Studies for Part II of the 'Philosophical Investigations', (eds. G. H. von Wright and H. Nyman, trans. C. G. Luckhardt and M. A. E. Aue), Oxford: Basil Blackwell 1982.
Z: *Zettel*, (eds. G. E. M. Anscombe and G. H. von Wright, trans. G. E. M. Anscombe), 2nd edition Oxford: Basil Blackwell, 1981.
OC: *On Certainty*, (eds. G. E. M. Anscombe and G. H. von Wright, trans. Denis Paul and G. E. M. Anscombe), Oxford: Basil Blackwell, 1977.

man.[2] Throughout the 18th and 19th centuries the machine furnished a central image or metaphor in terms of which attempts were made to understand man. This was true of the philosophy of that period in both its continental and more insular varieties, to both of which varieties of philosophy Wittgenstein was heir. The early Wittgenstein was not noticeably interested in matters psychological, but from the time of his return to philosophy in the late 1920s, he concerned himself with the task of elucidating the meaning and applicability of psychological terms. One recurrent aspect of that concern was manifested in his consideration of what it would mean to say that a machine thinks.[3] We believe that we know what a computer is and what it does and, in our more unreflective moments, we are also inclined to believe that we know what thinking is. On the basis of these naive beliefs it appears that whatever the answer to the question "Can computers think?" may be, the question itself is at root an empirical one, requiring a correspondingly empirical answer. Wittenstein, rejecting any simple univocal account of thinking, holds that that the question "Can computers think?" requires a conceptual rather than an empirical answer. In this article I want to consider what contribution, if any, Wittgenstein can make to the vexed question of whether a computer can think. We

---

[2] Rene Descartes, *Discourse on Method*, Part V, in Elizabeth S. Aldine and G. R. T. Ross (trans.), *The Philosophical Works of Descartes*, Volume I (Cambridge: Cambridge University Press, 1970). Descartes' conclusion, by the way, was negative:

> If there were a machine which bore a resemblance to our body and imitated our actions as far as it was morally possibly to do so, we should always have two very certain tests by which to recognize that, for all that, they were not real men. The first is, that they could never use speech or other signs as we do when placing our thoughts on record for the benefit of others. For we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for instance, if it is touched in a particular part it may exclaim
> that it is being hurt, and so on. But it never happens that it arranges its speech in various ways in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do. And the second difference is, that although machines can perform certain things as well as or perhaps better than any of us can do, they infallibly fall short in others, by the which means we may discover that they did not act from knowledge, but only from the disposition of their organs. For while reason is a universal instrument which can serve for all contingencies, these organs have need of some special adaptation for every particular action. From this it follows that it is morally impossible that there should be a sufficient diversity in any machine to allow it to act in all the events of life in the same way as our reason causes us to act. [p. 116]

We might consider this a proto-refutation of the Turing test! For a very interesting account of artificial intelligence which sees it as raising genuine metaphysical questions see Jose A. Benerdete, *Metaphysics: The Logical Approach* (Oxford: Oxford University Press, 1989).

[3] One of the founding fathers of modern artificial intelligence, Alan Turing, was an auditor of Wittgenstein's lectures on the foundations of mathematics and was powerfully affected by them. In turn, Wittgenstein had a copy of Turing's 1937 paper which he obviously had read since he commented on it. See M. Nedo and M. Ranchetti (eds.), *Ludwig Wittgenstein, Sein Leben in Bildern und Texten* (Frankfurt am Main, 1983), p. 309, cited in Otto Neumaier 'A Wittgensteinian View of Artificial Intelligence' in Rainer Born (ed.), *Artificial Intelligence: The Case Against* (London: Croom Helm, 1987), p. 133. See also RPP I, 1096.

know, or we believe that we know, what a computer is. (This may well be a more complicated matter than we think — but for the present I shall take it as being unproblematic.) We are also inclined to believe that we know what thinking is. On the basis of this belief it appears that whatever the answer to the question "Can computers think?" may be, the question is at root an empirical one. While I shall, in the end, defend this view, I hope to show, by means of an examination of Wittgenstein's ideas, that matters here are not quite as straightforward as they appear at first glance. I shall begin by giving a brief characterisation of artificial intelligence; then I shall present and discuss those passages in Wittgenstein's works explicitly concerned with "thinking machines" to determine his thought on the matter; next, I shall focus on Wittgenstein's account of thinking, and I shall conclude with an overall evaluation of Wittgenstein's position on artificial intelligence.

2. Artificial Intelligence

The term "Artificial Intelligence" was introduced to the world by John McCarthy and Marvin Minsky at a conference in Dartmouth, New Hampshire, in 1956. Whether or not it had a clear and unambiguous meaning at that time, since then it has acquired a range of interpretations. Here is a representative sample of some of those interpretations.

According to Alan Garnham[4] artificial intelligence is the science of thinking machines. It may be conceived of in two ways: 1. as being concerned with the production of useful machines, and 2. as an effort to understand human intelligence. Margaret Boden[5] believes that artificial intelligence is not concerned with the production of useful machines; it is, rather, concerned with the study of intelligence in thought and action. Computers are the tools of artificial intelligence, because its theories are expressed as computer programs that enable machines to do things that would requires intelligence if done by people. She writes

> By "artificial intelligence" I therefore mean the use of computer programs and programming techniques to cast light on the principles of intelligence in general and human thought in particular. In other words, I use the expression as a generic term to cover all machine research that is somehow relevant to human knowledge and psychology, irrespective of the declared motivation of the particular programmer concerned.[6]

---

[4] Alan Garnham, *Artificial Intelligence: An Introduction* (London: Routledge and Kegan Paul, 1988), p. xiii and p. 2.
[5] Margaret Boden, *Artificial Intelligence and Natural Man* (Hassocks, Sussex: The Harvester Press, 1977), Preface.
[6] Boden, 1977, p. 5.

Boden's understanding of the fundamental nature of artificial intelligence shows no basic change over the next eleven years, for in 1988 she writes "[The goal of) artificial intelligence...is to understand, whether for theoretical or technological purposes, how representational structures can generate behaviour and how intelligent behaviour can emerge out of unintelligent behaviour."[7]

Marvin Minsky believes that artificial intelligence is "the field of research concerned with making machines do things that people consider to require intelligence."[8] Robert Solso uses the term artificial intelligence "to embrace all forms of computer-produced output that would be considered 'intelligent' if produced by a human"[9] and Elaine Rich believes artificial intelligence to be "The study of how to make computers do things are which, at the moment, people are better."[10] According to Avron Barr and Edward Feigenbaum "Artificial Intelligence is the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behaviour — understanding language, learning, reasoning, solving problems, and so on".[11]

In these accounts from a variety of sources, there is solid agreement on one point: artificial intelligence has something to do with intelligent machines. Some of those concerned with artificial intelligence are more interest in intelligent machines, while others prefer to concentrate on intelligent machines. As we have seen from the citations above, a consensus has emerged as to what, in functional terms, the "intelligence" in "artificial intelligence" would be. This consensus is expressed in the form of a modal conditional, which runs, in Howard Gardner's formulation "artificial intelligence seeks to produce, on a computer, a pattern of output that would be considered intelligent if displayed by human beings."[12]

One could be committed to artificial intelligence in this sense without going as far as John Haugeland wants to go. He exhibits no inhibitions in claiming that "Artificial Intelligence [is] the exciting new effort to make computers think. The fundamental goal of this research is not merely to mimic intelligence or produce some clever fake. Not at

---

[7] Margaret Boden, *Computer Models of Minds* (Cambridge: Cambridge University Press, 1988), p. 6.
[8] Marvin Minsky, *The Society of Mind* (London: Picador, 1988), p. 326.
[9] Robert Solso, *Cognitive Psychology* (2nd edition Boston: Allyn & Bacon, 1988), p. 460.
[10] Elaine Rich, *Artificial Intelligence* (Singapore: McGraw-Hill Book Company, 1983), p. 1.
[11] Avron Barr and Edward Feigenbaum (eds.), *Handbook of Artificial Intelligence, Vol. I* (London: Pitman, 1981), p. 3.
[12] Howard Gardner, *The Mind's New Science: A History of the Cognitive Revolution* (New York: Basic Books, 1985), p. 140.

all. "AI" wants only the genuine article: machines with minds, in the full and literal sense."[13]

In order to distinguish among the different varieties of artificial intelligence and to see which of them, if any, could be of interest to philosophy, I shall adopt Owen Flanagan's taxonomy. Flanagan[14] postulates four different kinds of artificial intelligence. To begin with there is nonpsychological artificial intelligence. Here the artificial intelligence worker builds and programs computers to do things that, if done by us, would require intelligence. No claims are made about the psychological realism of the programs. As an esoteric branch of electronic engineering, artificial intelligence in this mode is of little or not interest to philosophers. In weak psychological artificial intelligence the computer is regarded as being a useful tool for the study of the human mind. Programs simulate alleged psychological processes in man and allow researchers to test their predictions about how these alleged processes work. This is the kind of artificial intelligence that J. Russell,[15] for example, takes to be relevant to cognitive psychology. Strong psychological artificial intelligence is the view that the computer is not merely an instrument for the study of mind but that it really is a mind (Haugeland's view). Finally, there is suprapsychological artificial intelligence which agrees with strong psychological artificial intelligence in claiming that mentality can be realised in computers but transcends the residual anthropological chauvinism of strong psychological in being interested in all the conceivable ways in which intelligence can be realised. Of these four kinds of artificial intelligence, only strong psychological artificial intelligence and suprapsychological artificial intelligence are of central interest to the philosopher.

3. Wittgenstein's Position on Artificial Intelligence

It has been claimed that Wittgenstein rejects outright the possibility of artificial intelligence.[16] It has also been claimed that he is a progenitor of artificial intelligence.[17]

---

[13] John Haugeland, *Artificial Intelligence: The Very Idea* (Cambridge, Mass.: The MIT Press, 1985), p. 2.
[14] Owen Flanagan, *The Science of Mind* (London: The MIT Press, 1984), 227- 230.
[15] J. Russell, *Explaining Mental Life: Some Philosophical Issues in Psychology* (London: Macmillan, 1984).
[16] H. L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence* (2nd revised edition new York: Harper & Row, 1979).
[17] Y. Wilks, 'Philosophy of Language' in E. Cherniak and Y. Wilks (eds.), *Computational Semantics: An Introduction to Artificial Intelligence and Natural Language Comprehension* (Amsterdam: 1976). The enterprise of discovering the ancestors of artificial intelligence has becomes something of an intellectual parlour game. Ernest Moody claims that Ramon Lull's invention of a machine with rotating discs for the purposes of automatic calculation "perhaps earns him the right to be called the father of computer programming". Ernest Moody, 'Medieval Logic' in Paul Edwards (ed.). *The Encyclopaedia of Philosophy*, Volume 4 (New York: Macmillan, 1967), p. 530. Gerard Casey and Aidan Moran identify Thomas Hobbes as the proto-ancestor of artificial intelligence because of his identification of ratiocination with computation—see G. N. Casey

But not even Wittgenstein can attack and defend the possibility of artificial intelligence at the same time and in the same respect! Let us consult the texts, beginning, in chronological order, with the *Blue Book* (1933-1934)

> [T]he problem here arises which could be expressed by the question: "Is it possible for a machine to think?" .... And the trouble which is expressed in this question is not really that we don't yet know a machine which could do the job. The question is not analogous to that which someone might have asked a hundred years ago: "Can a machine liquefy a gas?" The trouble is rather that the sentence "A machine thinks (perceives, wishes)": seems somehow nonsensical. It is as though we had asked "Has the number three a colour?" [BB, p. 47]

The problem with the sentence "Is it possible for a machine to think?" is that while grammatically it is a question, it is not really a question at all. Why not? A question is indicative of ignorance (real or feigned) on the part of the questioner. But the ignorance cannot be a total ignorance: a questioner cannot simply know nothing. Any question which really is a question operates within the horizon of a range of possible answers; the questioner has some idea (albeit inchoate) of what would constitute an answer to his question, otherwise he would not be in a position to recognise the answer when it came. For example, if I ask you now "What time is it?" and you respond "Oh to be in Ireland now that winter's here!" I should not consider that response an answer to my question. If, in a poetic vein, you had responded "Half past spring" then that response would just have been on the margins of acceptability as an answer to my question. If you had said "Two thirty" when the time was actually twelve thirty then your response, although incorrect, would undoubtedly have been an answer to my question. Every question, then, operates within boundaries that delimit what is to count as an answer and what is not. In the passage just cited Wittgenstein appears to be suggesting that the grammatical question "Is it possible for a machine to think?" is not really a question at all. It has no range of appropriate responses; we simply do not know what could count as a possible answer. Stuart Shanker, in his editorial introduction to Volume 4 of *Ludwig Wittgenstein: Critical Assessments*, makes essentially the same point — "It is not that computers lack consciousness, it is that the concept of consciousness simply cannot be applied to a machine".[18]

Next, let us consider a passage from the *Philosophical Grammar* (1933):

---

and A. Moran, 'The Computational Metaphor and Cognitive Psychology' in *Cognitive Science*, a special issue of the *Irish Journal of Psychology*, Vol. 10, no. 2, 1989, pp. 143-161.

[18] Stuart Shanker, *Ludwig Wittgenstein: Critical Assessments, Volume 4 (From Theology to Sociology: Wittgenstein's Impact on Contemporary Thought)* (London: Croom Helm, 1986), p. 12. See OC 314, 315 for Wittgenstein's remarks on legitimate questions.

> If one thinks of thought as something specifically human and organic, one is inclined to ask "could there be a prosthetic apparatus for thinking, an inorganic substitute for thought?" But if thinking consists only in writing or speaking, why shouldn't a machine do it? "Yes, but the machine doesn't know anything." Certainly, it is senseless to talk of a prosthetic substitute for seeing and hearing. We do talk of artificial feet, but not of artificial pains in the foot. "But could a machine think?" — Could it be in pain? — Here the important thing is what one means by something being in pain. I can look on another person — another person's body — as a machine which is in pain. And so, or course, I can in the case of my own body. On the other hand, the phenomenon of pain which I describe when I say something like "I have a toothache" doesn't presuppose a physical body. (I can have toothache without teeth.) And in this case there is no room for the machine — It is clear that the machine can only replace a physical body. And in the sense in which we can say of such a body that it is in pain, we can say it of a machine as well. or again, what we can compare with machines and call machines is the bodies we say are in pain.[19] [PG 64, p. 105]

In the *Remarks on the Foundations of Mathematics* (1942-1943) Wittgenstein asks "Does a calculating machine calculated [RFM IV §2] And in the Remarks on the Philosophy of Psychology I (1946- 1947) he remarks "Turing's 'Machines'. These machines are humans who calculate." [RPP 11096] Finally, from the *Philosophical Investigations*, we have the best known of Wittgenstein's passages on the topic of thinking machines:

> 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.[20] [PI 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.[21] [PI 360]

We may distil the essence of the foregoing citations into the following points. First, the question of whether a machine can think, insofar as it is a real question at all, is a

---

[19] In *Zettel* (1946-1948), in a passage reminiscent of the passage just cited from the *Philosophical Grammar*, Wittgenstein asks "Is thinking a specific organic process of the mind, so to speak — as it were chewing and digesting in the mind? Can we replace it by an inorganic process that fulfils the same end, as it were use a prosthetic apparatus for thinking? How should we have to imagine a prosthetic organ of thought?" [Z, 607]

[20] Stuart Shanker comments on this and similar passages: "The point Wittgenstein was raising was, however, solely concerned with the intelligibility of speaking of a mechanical calculation; with the question of whether it makes sense to describe the operations of such sophisticated machines as 'calculations', let alone as thinking, understanding, knowing, inferring, etc .... Calculation, as we understand the term, is a normative concept .... the concept of calculation, qua normative concept, demands the ability to follow a rule .... There is a distinction to be drawn between mechanical symbol-manipulation and (humanly effective) calculation; the philosophical problem we encounter is how to describe these machine operations, and how to elucidate the distinction between them and the correct application of a rule/algorithm". [Shanker, 1986, p. 11, p. 17]

[21] Cf. *Zettel* 614 "But must there be a physiological explanation here? Why don't we just leave explaining alone? — But you would never talk like that, if you were examining the behaviour of a machine! — Well, who says that a living creature, an animal body, is a machine in this sense?"

conceptual question, not an empirical one, and it can be answered, if it can be answered at all, only by conceptual analysis.[22] [BB, 47; PI 359] Second, machines can be likened to human beings only in respect of their bodies. [PG 64; PI 359] With these points in mind, Wittgenstein's argument can be summarised thus:

(i) We are entitled to predicate 'calculating', 'thinking', and 'knowing' only of human beings and of that which is sufficiently like a human being in the appropriate respects.
(ii) Machines are neither human nor sufficiently like human beings in the appropriate respects.
(iii) Therefore, we cannot predicate 'calculating', 'thinking', or 'knowing' of machines. [PG 64; RFM IV §2; RPP I 1096]

The argument is quite clearly valid; its soundness, however, is quite another matter, for while the first of the premises is something of a truism, the second premise is far from being self- evident. The soundness of Wittgenstein's argument, then, turns on the truth of this premise. To see what kind of evidence Wittgenstein is prepared to offer in its support I should like to examine his account of thinking for it is there, if anywhere, that we should find the relevant evidence.

4. Wittgenstein on Thinking

What is thinking? According to Wittgenstein, the concept of thinking is widely ramified, comprising many manifestations of life. [Z 110; RPP II, 218, 220, 234] He remarks "What a lot of things a man must do in order for us to say he *thinks*" [RPP I, 663] There is, indeed, no reason to expect the concept of thinking to have a unified employment — we should rather expect the opposite. [Z 112] It is Wittgenstein's belief that the use of 'thinking' is confused — as indeed is the use of all psychological verbs and the very science of psychology itself. [Z 113; RPP II 20, 194, PI II xiv, 232e; Z 462]

As a method of investigating the range of phenomena to which the term 'thinking' can be applied introspection is of little or no use. [RPP II 31, 35] You have as much chance as figuring out what 'think' means by watching yourself while you think, as you have of figuring out what the word 'checkmate' means by earnestly scrutinising the last

[22] Shanker comments "It just is unintelligible to apply a normative/intentional concept to a mechanical manipulation of a symbol .... If we argue that it is unintelligible to debate whether machines can think, the emphasis must be placed firmly on the unintelligibility of the mechanist thesis. In other words, the confusions which undermine the mechanist thesis are entirely conceptual, not empirical".[Shanker, 1986, p. 14; p. 9]

move in a game of chess.[23] [PI 316] Is thinking a mental process? According to Wittgenstein thinking can be said to be a process or activity of the mind but not in the same sense as writing is an activity of the hand. [PG p. 106] When we consider thinking as a form of activity, we consider it as a form of mental activity as distinct from a form of bodily activity. But a question arises over the presupposition that mental and physical activities are activities in exactly the same sense. [Z 123; cf. RPP II 193] Thinking can be called a mental process only if we are prepared to call seeing a written sentence or hearing a spoken sentence a mental process. In Wittgenstein's opinion, thinking can be called a mental process only if pain is a mental process, and calling either thinking or pain a mental process is intended to distinguish experience from physical processes. [PG p. 106]

In the context of a discussion of intention, Wittgenstein claims that what causes problems in regard to things mental is the very grammar of the word 'process', not the further question of what kind of process e.g. intention is. [PG p. 148] In *Zettel* we are counselled not to think of understanding as a mental process at all, for that way confusion lies. [Z 446; PI 158] Because we have a verb 'to understand' and because we believe that understanding is an activity of mind, we imagine that we must find a specific mental process underlying it. [Z 446]

To conceive of thinking as a process that goes on in secret is to risk being misled. [RPP I 580] Bodily processes, such as digestion, breathing, and so on, are strictly incomparable with so- called mental processes such as thinking, feeling, wanting. [RPP I 661] In the Investigations Wittgenstein claims that if we deny that thinking is an incorporeal process our denial does not derive from our intimate acquaintance with incorporeal processes and the resultant knowledge that thinking is not to be found among them. Such talk of processes comes from trying to explain the meaning of thinking in a primitive way. We use an expression such as 'incorporeal process' to distinguish the grammar of the word think from that of the word eat. [PI 339; cf. PI 154] We speak of understanding as a mental process, and the grammar of 'process' here is in many respects similar to the grammar of 'process' as it occurs in 'Brain process'. However, Wittgenstein notes that there is this salient difference between the two: in the

[23] According to C. Grant Luckhardt "Wittgenstein also denies that "inner" nonphysical processes can serve as the objects to which psychological predicates refer .... His objection is not that they would be meaningful only to their authors, but that it is not clear that they would have any meaning, even to their authors .... Privately introspectible objects, according to Wittgenstein, just won't do the job they are intended to do — justifying the application of psychological verbs to oneself. 'Wittgenstein and Behaviourism', *Synthese* 56 (1983), p. 327.

case of brain process, a direct check is possible in principle; in the case of mental process, no such direct check is possible. [PG 41, p. 83] It is important to be quite clear that Wittgenstein is not denying the reality of mental process. What he is doing is setting his face against the picture of an inner process, against the view that this picture of an inner process gives us the correct idea of the use of psychological verbs.

> How does the philosophical problem about mental processes 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 this 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. (The decisive movement in the conjuring trick has been made, and it was the very one we thought quite innocent.) — And now the analogy which was to make us understand uncomprehended processes falls to pieces. So we have to deny the yet uncomprehended process in the unexplored medium. And now it looks as if we had denied mental processes. And naturally we don't want to deny them. [PI 308]

The passage just cited raises the spectre of behaviourism, a malady with which many suspect Wittgenstein is afflicted.[24] Sometimes, it seems as if Wittgenstein even suspects himself! In the passage immediately prior to the one just cited the irrepressible interlocutor asks "Are you not really a behaviourist in disguise? Aren't you at bottom saying that everything except human behaviour is a fiction?" But Wittgenstein denies that he is a behaviourist. "If I do speak of a fiction, then it is of a grammatical fiction." [PI 307] In the *Philosophical Grammar* Wittgenstein says that understanding is not to be identified with the behaviour that shows us the understanding. Understanding is a state of which behaviour is a sign. [PG §41, p. 84] In *Zettel* he remarks that "Joy is not joyful behaviour". [Z 487] In the Investigations he notes that a question could be asked as to whether psychologists study behaviour and not the mind. Wittgenstein's answer to that question is that psychologists observe the behaviour of human beings, particularly their verbal utterances, but it is not the case that theses utterances are about behaviour. [PI II v 179e] And in a very striking passage from *Remarks on the Philosophy of Psychology II* Wittgenstein is perfectly explicit is his refusal to equate thinking with behaviour.

> The word "thinking" is used in a certain way very differently from, for example, "to be in pain", "to be sad", etc.: we don't say "I think" as the expression of a mental state. At most we say "I'm thinking". "Leave me alone, I'm thinking about . . ." And of course one

[24] See Richard Rorty, 'Wittgensteinian Philosophy and Empirical Psychology', *Philosophical Studies* 31 (1977), p. 169; Robert J. Fogelin, *Wittgenstein* (London: Routledge and Kegan Paul, 1976), p. 176; Arnold S. Kaufman, 'Behaviourism', in The *Encyclopaedia of Philosophy*, Paul Edwards (ed.), (New York: The Macmillan Company, 1967), Vol. I, p. 270. Cathal Daly, 'New Light on Wittgenstein: Part Two', *Philosophical Studies* (Irl.) XI (1961-1962).

> doesn't mean by this: "Leave me alone, I am now behaving in such and such a way." Therefore "thinking" is not behaviour. [RPP II, 12]

In view of Wittgenstein's own claim that he is not a behaviourist and the critical acceptance of that claim by commentators as long ago as 1961, it is odd that more recent critics should have found it necessary to argue the point.[25] Is there something in Wittgenstein's writings which could substantiate a counter claim? Well, for one thing, Wittgenstein is unrelenting in his attack on the relics of Cartesian dualism. If one accept the Cartesian problematic and if one rejects the Cartesian spiritual substance (soul) then one is left with the Cartesian material substance (body) and its behavioural manifestations. Another aspect of Wittgenstein's thought which might lead to his being considered a behaviourist is his clear recognition that the subjects which a psychologist studies are not available for his inspection in exactly the same way as are the subject which a physicist studies. Seeing, hearing, thinking, feeling, willing — all these are observed by the psychologist through the external reactions (the behaviour) of the subject. [PI 571] Despite these indications to the contrary, there is really no difficulty in taking Wittgenstein at his word. Luckhardt, for example, suggests that while the language game of mental states is played, or understood, through the language game of behaviourism, the relation between these two language games is not one either of entailment or of causality.[26] In Luckhardt's view, playing the mental states language game through the behaviour language game requires skills of interpretation, the ability to see things as beings of certain kinds.

Of the theories of mind currently enjoying popularity, the one most compatible with the view that machines can think is the Identity Theory, according to which mind and brain are identical. The goal of strong artificial intelligence, machines with minds in the full and literal sense, would seem to be eminently compatible with the Identity Theory. If it could be shown that Wittgenstein, having rejected Cartesian Dualism, must embrace the Identity Theory then, given that theory's compatibility with the claims of strong artificial intelligence, the claims of those who see in Wittgenstein a supporter of the claims of strong artificial intelligence would be indirectly supported. Is Wittgenstein, then, an Identity Theorist?

---

[25] Luckhardt 1983 and Neumaier 1987.

[26] Luckhardt 1983, p. 333. This is very close to Neumaier's distinction of two senses of the term 'criteria' as used by Wittgenstein; a strict entailment sense in which criteria are necessary and sufficient conditions of the presence of that of which they are the criteria, and a looser sense in which criteria do not guarantee the presence of that of which they are the criteria. Behaviour, as criterion, is a criterion only in this looser sense. It is never a strict logical guarantor of the existence or nature of any given mental state. See Neumaier 1987, pp. 142-146.

In the *Blue Book* Wittgenstein notes that there are two kinds of proposition: those describing facts, and those describing personal experiences. Consequently, there appear to be two worlds corresponding to these kinds of proposition, a physical world and a mental world. The mental world we are apt to think of as being gaseous or aethereal, but this idea of an aethereal object is a subterfuge which arises simply from our embarrassment with the grammar of certain words. All that we are really affirming here is the negative claim that these words are not being used as names for material objects. [BB p. 47] In *Remarks on the Philosophy of Psychology I*, Wittgenstein goes further:

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought- processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thought. But why should the system continue further in the direction of the centre. Why should this order not proceed, so to speak, out of chaos? The case would be like the following — certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced — but nothing in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that comes out of it — this can only be done from the history of the seed. So an organism might come into being even out of something quite amorphous, as it were causelessly; and there is no reason why this should not really hold for our thoughts, and hence for our talking and writing. [Z 608; RPP I 903]

To refute the notion that there must be a physiological basis for our mental activities, Wittgenstein suggests the following thought experiment. We want someone to remember a text. While we read it to him, he scribbles on paper. But the marks he makes are not writing, not a translation of what we are saying into another symbolism. Yet without these doodles, he is unable to repeat the text. In effect, the doodles are not a stored up version of the text. Why then, Wittgenstein asks, should we consider that our mental activities are stored up in the central nervous system?[27] [Z 612]

If Wittgenstein is hostile to the Identity Theory,[28] he is, as we have seen, no less hostile to its major modern rival Psychophysical Parallelism. In Wittgenstein's view, both the Identity Theory and Psychophysical Parallelism are responses to a fundamentally

---

[27] See *Last Writings on the Philosophy of Psychology* 922-924. In *Zettel* Wittgenstein suggest that it might not be possible to investigate certain psychological phenomena physiologically for the good and sufficient reason there is nothing physiological corresponding to them! [Z 609; RPP I, 904]

[28] Norman Malcolm, perhaps with tongue in cheek, suggests a connection between the Identity Theory and the occult! "Could it be that the conception of mind as brain — a view that is supposed to be 'scientifically oriented' — is actually nourished by the charm of the occult? I think so. Popular lectures and articles speak of the brain with particular awe. What a marvellous mechanism it is! Certainly the brain is marvellous — but so too is the heart and the digestive system. Why should the brain attract special awe? Surely because it is conceived of as 'the organ of thinking'." Norman Malcolm, *Nothing is Hidden: Wittgenstein's Criticism of his Early Thought* (Oxford: Basil Blackwell, 1986), p. 200.

mistaken problematic. Psychophysical Parallelism comes into being as a result of a primitive interpretation of our concepts, for a denial of a physiological mediation between psychological phenomena seems to commit one to a belief in a gaseous mental entity, a soul alongside a body. [Z 611; RPP I 906] The truth is, however, that the feeling of an unbridgeable gulf between consciousness and brain, a feeling which to a large extent fuels all forms of dualism, is not something of which one is ordinarily aware. It occurs only when one turn one's attention to one's own consciousness; and this is a very queer thing, this turning of attention to one's own consciousness. [PI 412]

5. Critical Comments

We have seen from the passages cited above that Wittgenstein is not willing to attach psychological attributes to machines because of the manifest and significant dissimilarities between machines and human beings.[29] The evidence to support

[29] Despite the obvious intent of these passages, Klaus Obermeier claims that "a Wittgensteinian account of 'understanding' does indeed support claims made by advocates of 'strong AI' that pertain to the role and function of language in the understanding process." Klaus Obermeier, 'Wittgenstein on Language and Artificial Intelligence: The Chinese-Room Thought Experiment Revisited', *Synthese* 56 (1983), p. 340. How can Obermeier maintain this? He claims that a distinction can be drawn, based on the thought of Wittgenstein, between understanding-as-performance (UP) and understanding-as-feeling (UF). According to Obermeier, many objections to strong artificial intelligence are based on the conflation of these two distinct kinds of understanding. Computer programs are capable of UP and hence can be said to be intelligent. Obermeier's argument (a complex and difficult one, only sketchily represented here) contains a rather brusque dismissal of the relevance of intentionality to the problem of thinking machines. This dismissal should be seen against the background of the principled objections to strong artificial intelligence expressed by, among others, Haugeland and Boden, which have their roots in the notion of intentionality. I should make clear, before I proceed, that these objections, thought expressed in the works of Haugeland and Boden are not necessarily representative of their own final positions.

Boden (1988) points out that "controversy attends the question whether computational psychology can in principle explain the higher mental processes" (p. 8) She adds "most computational psychologists trust that their approach will explain how representations function, and many believe it will even help to illuminate what representations are. But a few . . . argue that computer models (and psychological theories grounded in them) in principle cannot exhibit or explain genuine representation (or meaning) at all", (p. 8) The heart of the problem resides in the computationalist's identification of the mental processes with computation. [Boden 1988, p. 229] This identification has an ancient philosophical lineage, its prototypical exponent being Thomas Hobbes who claims that "REASON...is nothing but reckoning" [Thomas Hobbes, *The Leviathan* (Sir William Molesworth ed., London: J. Bohn, 1839-1845, Vol. Ill, p. 30 Originally published in 1651.] and "By RATIOCINATION, I mean computation" [Thomas Hobbes, *The Elements of Philosophy* (Sir William Molesworth ed., London: J. Bohn, 1839-1845, Vol. I, p. 3.]

As Haugeland (1985) points out, in Hobbes's view, thinking consists of symbolic operations in which thoughts are not spoken or written symbols but special brain tokens. We see than that the central assumptions of cognitive science are essentially the same as those of Hobbes's on the nature of reason. According to Pinker & Mehler, the central assumption of cognitive science is that "intelligence is the result of the manipulation of structured symbolic expressions". [S. Pinker & A. J. Mehler, "Introduction," *Cognition* 28, 1988, pp. 1- 2; cf. S. Pinker and A. Prince, 'On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition,' *Cognition* 28, 1988, pp. 73-193.] Haugeland states that "cognitive science rests on a profound and distinct empirical hypothesis: that all intelligence, human or otherwise, is realised in rational, quasi-linguistic symbol manipulation." [Haugeland 1985, pp. 249-250] and Boden claims that computational psychology "covers those theories which hold that mental processes are . . . the sorts of formal computation that are studied in traditional computer science and symbolic logic." [Boden 1988, p. 229]

Wittgenstein's position is to be found in his remarks on thinking which I have just examined. In Wittgenstein's view, as we ordinarily use the terms which signify psychological characteristics, it simply makes no sense to attribute them to machines. We cannot, while claiming to be coherent, speak of machines as thinking, calculating, having intentions, and so on. Now, as a statement of fact, this seems to be true, thought perhaps not quite as clearly so as it was 40 years ago. But is this lack of sense in the attribution of psychological predicates to machines simply a matter of current verbal usage reflecting the current way things are; a usage which is, in principle, revisable so that we might, conceivably, someday have a language game in which the attribution of such predicates to machines might be in order? If this is all that Wittgenstein intends then his argument is true but is hardly of any great significance. If the argument is to be significant, then verbal usage must reflect not simply the way things are, but also the way things must be.

In short, Wittgenstein's argument against artificial intelligence, if it is to be philosophically significant, requires as a second premise not one which holds that machines as a matter of fact do not resemble human beings, etc. but one which holds that in principle they cannot do so. Are the considerations which Wittgenstein brings to bear on the notion of thinking and other psychological activities sufficient to support the

---

There is a fundamental difficulty with this most basic assumption of the computational approach to cognition, a difficulty which is pithily expressed by Haugeland: "Hobbes . . . cannot tell the difference between minds and books. This is the tip of an enormous iceberg that deserves close attention, for it is profoundly relevant to the eventual plausibility of Artificial Intelligence. The basic question is: How can thought parcels mean anything?" [Haugeland 1985, p. 25] Haugeland calls this difficulty "the mystery of original meaning", the point of this phrase being that once meaning enters a system it can be processed in various ways, but the crucial question is how it got into the system in the first place. Hobbes and his latter-day disciples appear to have no answer to this question. Haugeland himself devotes a lot of space in his book to this topic but he too is unable to arrive at a satisfactory resolution. An essentially similar point has been made by John Searle in his notorious "Chinese Room" thought experiment. Unlike some other critics of the computational model, Searle is willing to allow that machines can encompass the feat of generating original meanings, but only if they are biological machines. It is only fair to point out that controversy rages in the philosophical journals on the merits and demerits of Searle's thought experiment, and gallant attempts have been, and are being, made to show how non-biological physical symbols systems can embody intentionality. See, for example, D. Anderson, 'Is the Chinese Room the Real Thing?' *Philosophy* 62, 1987, pp. 389-393; T. W. Bynum, 'Artificial Intelligence, Biology and Intentional States', *Metaphilosophy* 16 1985, pp. 355-377; L. R. Carleton, 'Programs, Language Understanding and Searle', *Synthese* 59, 1984, pp. 219-230; and R. Lind, 'The Priority of Attention: Intentionality for Automata", *Monist* 69, 1986, pp. 609-619.

A similar difficulty arises with the related notion of 'information'. Boden asks "But what is "information"? Doesn't it have something to do with meaning, and with understanding? Can a computer mean, or understand — or even represent — anything at all?" [Boden 1988, p. 225] Westcott claims that "psychologists forgot that the notion of "information" as developed by Shannon . . . was absolutely meaningless. Information is merely a measure of channel capacity, admittedly important to communications theory; but "information" bears not significance other than its occupancy of this channel capacity". [M. R. Westcott, 'Minds, Machines, Models, and Metaphors: A Commentary,' *The Journal of Mind and Behaviour* 8, 1987, p. 287.] Similarly, Bakan claims that "The defect of the scientific universe of discourse is that it has no place in the objective world for information, except information in the bound [i.e. materially embodied] condition." D. Bakan, 'On the Effect of Mind on Matter,' in R. W. Rieber (ed.), *Body and Mind* (New York: Academic Press, 1980), p. 18. italics in original.

modal version of the second premise required to make the argument philosophically interesting?

For Wittgenstein words have meaning and application only within forms of life or as parts of a language game. But language games just are what they are, and if they change, there is no rationale for their changing. Wittgenstein states baldly "Instinct comes first, reasoning second. Not until there is a language-game are there reasons."[30] In *On Certainty* Wittgenstein's final thoughts are expressed thus:

> I want to regard man here as an animal; as a primitive being to which one grants instinct but not ratiocination. As a creature in a primitive state. Any logic good enough for a primitive means of communication needs no apology from us. language did not emerge from some kind of ratiocination . . . You must bear in mind that the language game is so to say something unpredictable. I mean: it is not based on grounds. It is not reasonable (or unreasonable). It is there — like our life. [OC 475, 559]

So, unless there is a meta-language game, of which other language games are parts, there cannot be any reason from moving from one language game to another any more than there can be a reason for not moving from one language game to another. If a language game were to arise, who knows how, in which one committed no grammatical solecism by attributing thinking to machines, then machines in the context of that language game could be said to think!

This raises a very fundamental and not entirely novel question of the chicken and egg variety: which comes first, language or the world? Or, to rephrase the question, could either element in the pair (language-world) exist without the other and, if so, which one? The answer to me is obvious. What is — reality in some basic sense — is, and must be, prior to any finite language. It seems to me to be simply incontrovertible that reality in its fundamental aspects cannot be constituted by human thought or speech.[31]

To believe that human language and thought is constitutive of reality in some fundamental way would be to attribute divine characteristics to man which he simply does not manifest, either individually or collectively. Surely, whether or not machines will ever be able to think in the way in which we are able to think cannot simply be a matter of what we can or cannot say at a given time, *unless what we can say is fundamentally dependent*

---

[30] *Remarks on the Philosophy of Psychology II* 689; Cf. *Philosophical Investigations* 371; 373, and *Remarks on the Philosophy of Psychology II* 632. See also R. G. Collingwood on change in what he calls "absolute presuppositions" in *Metaphysics* (Oxford: Clarendon Press, 1940), p. 48, n.; Stephan Koerner, *Categorial Frameworks* (Oxford: Basil Blackwell, 1974), Chapter VI, and Gerard Casey, *Natural Reason* (New York: Peter Lange, 1984), pp. 254-303.

[31] I stress 'fundamental' for certain aspects of reality, particularly its social or political or cultural or artistic dimension, are at least partially constituted by human thought and language.

upon the way things are. Henry Veatch gives the name "The Fallacy of Inverted Intentionality" to that position which accords priority to second intentions over first intentions:

> It is not the impossibility of a thing's being red and green at the same time that is determined by the rule [namely, that we cannot use colour words in this way], but rather the rule about "red" and "green" that is determined with a view to the use of these expression to signify just that very impossibility. Take away the possibility, and what could be the point of the rule?[32]

I agree with Veatch. If our language games make it senseless to predicate psychological terms of machines, then either it is an arbitrary and capricious piece of 'machinism' on our part, or else it is a reflection of what we take to be the nature of things.

Where does Wittgenstein stand on the question of the relative priority of language and world? I think it is true to say that although he is difficult to pin down on this matter, in the end he gives priority to language. If this is so, however, then the senselessness of otherwise of making certain attributions becomes inescapably indexical. It will not be categorically impossible to say "It is senseless to predicate psychological terms of machines" but only to say "It is senseless, in the Ordinary English Language (Cambridge Variety), 1951, to predicate psychological terms of machines".

Who can set boundaries to the march of a language game? Language games list where they will, without limitation, without reason. If there is no ontological status quo ready to provide a recalcitrant resistance to the vagaries of speech patterns, there is nothing to prevent the devotees of artificial intelligence from developing a language game in which psychological and intentional concepts are predicated univocally of man and machines.

hat there is a possibility of such a conceptual coup d'etat resulting in the installation of a new conceptual junta has dawned on some Wittgensteinians. Shanker has adverted to the possibility, but he is concerned not so much with the possible elevation of machine operations to the level of human mentality as with the reduction of human mentality the level of machine operations!

> The obvious worry is that, if you institute a conceptual revolution in the concept of thought so that it henceforth becomes intelligible to describe mechanical operations as thinking, then conversely there seems little reason why the argument should nor proceed in the opposite direction, thereby denying human beings the notions of autonomy and consciousness which underpin our conception of man as a rule-following creature.[33]

---

[32] Henry Veatch, *Two Logics* (Evanston, Illinois: Northwestern University Press, 1969), p. 121.
[33] Shanker 1986, p. 25.

It would seem then that the question "Can machines think?" cannot be, *pace* Wittgenstein, simply a conceptual question, something that can be determined by linguistic fiat. It must be a real question (in the sense on which I tried to distinguish real from pseudo questions above) which is resolvable in the context of an adequate ontology. Wittgenstein's argument is sound if directed against the actuality of artificial intelligence; if directed against the possibility of artificial intelligence then it must either arbitrarily denounce AI-compatible language games or suffer itself to be outflanked by the seemingly inevitable and, if Wittgenstein is to be believed, arational change of one language game for another.