**The "Hercules" in the Machine**
**Why Block's Argument Against Behaviorism is Unsound**[1]

Hanoch Ben-Yami

ABSTRACT *Ned Block argued, in his 'Psychologism and Behaviorism', that a behaviorist conception of intelligence is mistaken, and that the nature of an agent's internal processes is relevant for determining whether the agent has intelligence. He did that by describing a machine which lacks intelligence, yet can answer questions put to it as an intelligent person would. The nature of his machine's internal processes, he concluded, is relevant for determining that it lacks intelligence. I argue against Block that it is not the nature of its processes but of its linguistic behavior that is responsible for his machine's lack of intelligence. As I show, not only has Block failed to establish that the nature of internal processes is conceptually relevant for psychology, in fact his machine example actually supports some version of behaviorism. As Wittgenstein has maintained, as far as psychology is concerned, there may be chaos inside.*

1. *Cyrano de Bergerac*, Act III. Night. Roxane at her window, Christian before her balcony, Cyrano underneath it.

Roxane
You do not love me any more
Christian (*to whom* Cyrano *whispers his words*)
No– No– Not any more– I love you… evermore… and ever… more and more!
Roxane (*about to close her window—pauses*)
A little better…
Christian (*same business*)
Love grows and struggles like… an angry child… breaking my heart… his cradle…
Roxane (*coming out on the balcony*)
Better still—But… such a babe is dangerous; why not have smothered it new-born?
Christian (*same business*)
And so I do… And yet he lives… I found… as you shall find… this new-born babe… an infant… Hercules!
Roxane (*further forward*)
Good!–
Christian (*same business*)

---

[1] This is a revised version of my 'Behaviorism and Psychologism', first published in *Philosophical Psychology*, **18**(2), 2005, 179-86. Apart from some minor changes, I have added an appendix in which I reply to some objections made by Ned Block to my original paper. I have also restored the paper's original title, which the journal's editor did not like for some reason (the subtitle is unchanged).

> Strong enough… at birth… to strangle those two serpents—Pride and… Doubt.

2. What is the relevance of processes occurring within an agent to the agent's psychology? These processes certainly have a *causal* role in our psychology: what happens in our nerves and muscles causes our movements and produces speech, and our psychology is at least expressed in what we do and say. This causal role is, of course, one of the main factors that make the study of the internal structures and processes involved in our behavior of much interest—a study carried out mainly by neuroscience and other branches of physiology.

But do these internal processes have any analogous *conceptual* role as well? That is, are they *conceptually* linked to agents' psychology? When we use psychological predicates to characterize an agent, are we committed, by virtue of the *meaning* of these predicates, to anything concerning the nature of the agent's internal processes?

Behaviorism maintains that we are not. Internal processes' role is causal, but not conceptual—apart, perhaps, from the requirement that our movements and speech be produced by internal, and not external, causes. According to behaviorism, as I shall be using the term hereafter, what matters for a person's psychology is what that person does and would do in various circumstances. Following Wittgenstein's use of 'criteria', I shall say that psychological concepts have behavioral criteria.[2] According to behaviorism, no matter to what extent two persons differ in their internal processes, if these differences would not show in their behavior, then these persons are psychologically the same.

Indeed, Wittgenstein's own opinion also was that internal processes are conceptually irrelevant to psychology. In his *Zettel* (§§608-9) he went as far as to maintain:

> 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 thoughts. But why should the *system* continue further in the direction of the center? Why should this order not proceed, so to speak, out of chaos?
>
> It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them.

---

[2] Wittgenstein himself used this terminology with respect to psychological concepts or phenomena in, for instance, the *Philosophical Investigations*, § 580: 'An "inner process" stands in need of outer criteria.' Although I think criteria behaviorism is close to Wittgenstein's own position in some respects, the exact relation between the two positions will not be considered in this paper. For that purpose I would have to develop criteria behaviorism in some detail, yet this is not my purpose here. My purpose is, rather, to show that one influential criticism of those sorts of theories is invalid. Still, I do share with Wittgenstein, as the next paragraph will make clear, a negative appreciation of inner processes' conceptual role in psychology.

If no internal process may correspond to thinking, or to any other psychological phenomenon, then psychological concepts do not imply anything about the nature of internal processes. Internal processes have, in this case, no conceptual role in psychology.

Most philosophers nowadays believe that behaviorism is false. Internal processes *are* linked, they maintain, and linked *conceptually*, to psychology. Two persons may have the same behavioral dispositions, capacities and all other behavioral characteristics, yet be different psychologically, the difference being due to differences in their internal processes. For instance, one of these persons may be intelligent, while the other lacks any form of intelligence.

Several arguments raised against the *reductive* behaviorism of the forties and fifties brought about its current justified and prevailing rejection. I think, however, that these arguments are ineffective against behaviorism as formulated above, i.e., *criteria* behaviorism. My purpose in this paper is to defend this criteria behaviorism against one of its more influential criticisms, that of Ned Block in his paper 'Psychologism and Behaviorism' (Block, 1981; all page references to Block are to this paper).

'The doctrine that whether behavior is intelligent behavior depends on the character of the internal information processing that produces it' is called 'psychologism' by Block (p. 5). In his paper, Block argues that psychologism is true. His argument is considered by many to be one of the strongest arguments against behaviorism.[3] Block himself argues convincingly in his paper that what he takes to be "the main objections to behaviorism" already existing in the literature (p. 12, fn. 8) are inadequate to defeat a behaviorist conception of intelligence (p. 5). Thus, if Block's own argument can be shown to be ineffective, behaviorism, in the sense meant here, should be reconsidered: perhaps Wittgenstein was right after all, and as far as our psychological concepts are concerned, there may be chaos inside.

## 3.

Block argues against behaviorism by describing a machine that gives sensible answers to questions put to it, although it lacks any form of intelligence. Block maintains that his machine should be considered intelligent according to a behaviorist conception of intelligence, yet knowledge of the machine's internal information processing shows that it is not. I agree with Block that the machine he describes lacks intelligence. But I shall argue against him that, firstly, it is not the nature of his machine's internal information processing which is responsible, from a conceptual point of view, for its being devoid of intelligence; and, secondly, that his machine lacks intelligence according to criteria behaviorism as well.

In accordance with the classical Turing Test, Block first describes a machine that is limited to typewritten inputs and outputs, and that would pass the test although it lacks intelligence. He then generalizes his example by describing a robot that lacks intelligence yet acts in every possible situation like an intelligent person (pp. 23-4). His robot is an application of the principle according to which his machine was constructed (to be described below) to any form of sensory input and behavioral output. A criticism of the conclusions Block draws from his machine would therefore apply, with

---

[3] See, for instance, Braddon-Mitchell & Jackson (1996, pp. 111-20), who use only Block's argument to refute the behaviorist doctrine that "*all* that matters for having a mind is being such as to ensure the right connection between inputs and outputs" (p. 111). (Braddon-Mitchell & Jackson do not consider this doctrine behaviorist, but this need not concern us here.)

appropriate changes, to those he draws from his robot. For simplicity's sake I shall therefore describe and discuss only Block's machine.

This machine is programmed by a team of programmers. I reproduce, with some omissions, Block's description of it:

> First, we require some terminology. Call a string of sentences whose members can be typed by a human typist one after another in an hour or less, a *typable* string of sentences. Consider the set of all typable strings of sentences. Since English has a finite number of words …, this set has a very large, but nonetheless finite, number of members. Consider the subset of this set which contains all and only those strings which are naturally interpretable as conversations in which at least one party's contribution is sensible… Call a string which can be understood in this way a *sensible* string. For example, if we allot each party to a conversation one sentence per "turn" (a simplification I will continue to use), and if each even-numbered sentence in the string is a reasonable conversational contribution, then the string is a sensible one…
>
> Imagine the set of sensible strings recorded on tape and deployed by a very simple machine as follows. The interrogator types in sentence *A*. The machine searches its list of sensible strings, picking out those that begin with *A*. It then picks one of these *A*-initial strings at random, and types out its second sentence, call it "*B*." The interrogator types in sentence *C*. The machine searches its list, isolating the strings that start with *A* followed by *B* followed by *C*. It picks one of these *ABC*-initial strings and types out its fourth sentence, and so on…. (pp. 19-20)

Because of the way it was programmed, the machine emits only 'sensible' strings, strings in which *its* contribution is always sensible. It would, therefore, pass any test that is limited to conversational output and which determines that a conversant is intelligent only according to behavioral criteria—linguistic criteria, in this case. Block terms such a test a 'neo-Turing Test'. As he writes, a neo-Turing Test is one which determines that respondents have conversational intelligence if they have "the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be" (p. 18). However, it seems clear that Block's machine lacks intelligence:

> So long as the programmers have done their job properly, such a machine will have the capacity to emit a sensible sequence of verbal outputs, whatever the verbal inputs, and hence it is intelligent according to the neo-Turing Test conception of intelligence. But actually, the machine has the intelligence of a toaster. *All the intelligence it exhibits is that of its programmers*…. (p. 21)

Block has shown that despite the fact that his machine passes the neo-Turing Test, it lacks intelligence. The claim that passing the neo-Turing Test is sufficient for being intelligent has thus been refuted. Now Block's neo-Turing Test is behaviorist in nature. Its refutation is consequently taken by Block as a refutation of a behaviorist conception of intelligence:

> I conclude that the capacity to emit sensible responses is *not* sufficient for

> intelligence, and so the neo-Turing Test conception of intelligence is refuted… I also conclude that whether behavior is intelligent behavior is in part a matter of how it is produced. Even if a system has the actual and potential behavior characteristic of an intelligent being, if its internal processes are like those of the machine described, it is not intelligent. So psychologism is true. (p. 21)

Block's example was supposed to show that two beings can have the same "actual and potential behavior," while only one of them is intelligent. If so, then, since intelligence is a psychological phenomenon, the criteria of psychological concepts are not merely behavioral. Moreover, that Block's machine lacks intelligence was supposed to be determined by the nature of its internal processes. If so, then internal processes do have a conceptual role in an agent's psychology.

4. Block's machine may have reminded readers of Searle's Chinese Room (Searle, 1980). However, despite some similarities, the two thought-experiments are different in both aim and structure. Before proceeding to a criticism of Block's argument, I shall digress to a short discussion of these experiments' relation. I shall assume acquaintance with Searle's thought-experiment.

Firstly, Searle's thought-experiment is intended to show that *no* formal symbol manipulation of the kind carried out by computers would make a machine intelligent, even if the machine passes Turing's test . Block's, by contrast, was intended to show that *some* formal symbol manipulations that pass that test would not make a machine intelligent. Block's position, but not Searle's, is compatible with the view the some formal symbol manipulations that pass Turing's test *would* make machines intelligent. Indeed, Block has defended this position against Searle's argument (Block, 1980).

Block's thought-experiment was intended to support a position that was becoming orthodoxy in the philosophy of mind: functionalism. Searle's, by contrast, was intended to refute that same position, at least in one of its prevailing versions (see Putnam, 1975, essays 18, 20 and 21). For that reason Searle's thought-experiment attracted much discussion and criticism, many trying to show where it had gone wrong; while Block's was welcomed as establishing what most believed *must* be true.

Secondly, because of their different purposes, Block describes in detail the programming of his machine, while Searle does not describe any program (what Searle does describe is the way in which the person in the room would execute *any* program).

Lastly, the two philosophers also claim to rely on different kinds of intuition in their thought-experiments. Block wants us to realize that the programming of his machine is insufficient for intelligence. Searle, by contrast, wants us to see that since the person executing the program does not understand Chinese, a machine that would execute the same program would not understand Chinese either. The kinds of lack of intelligence or understanding that both authors want us to imagine are different. Block, in fact, argued against the Chinese Room's supposed intuition in his mentioned reply to Searle.

Accordingly, despite the fact that both Block and Searle describe scenarios in which a program that passes Turing's test lacks understanding or intelligence, their arguments are considerably different. Each deserves a separate criticism or evaluation. I shall now proceed to a criticism of Block's.

5. As previously mentioned, I agree with Block that his machine lacks

intelligence. However, contrary to Block, this is *not* determined by reference to its internal processes. The machine lacks intelligence because all it does is reproduce answers that were given to it in advance. And this is determined by reference to the relation between the answers that were formerly given to it and to the answers it now gives. The machine is not intelligent for the same reason that Christian is not a poet: Christian answers Roxane what Cyrano tells him to answer, and the machine answers its interrogator what the programmers 'told' *it* to answer.

Of course, our knowledge of the internal processing of Block's machine helps us to determine that it lacks intelligence. But it does that because we infer, relying on this knowledge, that Block's machine reproduces answers that were given to it by someone else. That is why this knowledge may *appear* relevant. But it is not conceptually relevant *in itself*—what *is* conceptually relevant is the conclusion that it enables us to draw about the nature of the machine's answers. One can indeed say that it is *in virtue of* the machine's internal processes that it exhibits the intelligence of its programmers—the italicized words here signifying the internal processes' *causal* role, that of mediating between programmers and machine. But it is *in virtue of* its merely reproducing the programmers' answers that it lacks intelligence—and now these words signify this fact's *conceptual* role.

The conceptual irrelevance of the nature of the machine's internal processes is demonstrated by the fact that other machines with different internal processes but with the same relation between linguistic input and output also lack intelligence. Block himself describes a variant of his machine, which can give the same answers but differs in the way it searches for them, and also lacks intelligence (p. 20). And both machines lack intelligence for the same reason: they give the answers that they were programmed to give. And other variants of Block's machine can also be imagined, in which, although the internal information processing is again different, the relation between linguistic input and output is the same or similar, and consequently these machines are also devoid of intelligence.

In fact, if it were an empirical discovery, that a machine gives as answers only what has been typed for it in the past as responses for the questions now put to it, and were we *entirely ignorant* of its internal processes, we would still be justified in denying it any intelligence.

Analogously, we do not know or need to know how Christian is processing what he hears, in order to determine that he is no poet; all we need to know is that he repeats whatever Cyrano says. Standing there, passionate, in the shadows of the night, before Roxane's balcony, there might very well be chaos inside his head; and, given what we need to know, the machine's answers may also proceed out of chaos; this is of no conceptual relevance. What is relevant is that both respond with answers given to them by someone else.

But does the fact that Cyrano supplies Christian with responses in real time, while the programmers have supplied the machine with all possible responses in advance, make any difference as regards the question whether the machine, but not Christian, passes Block's neo-Turing Test? In that case the Cyrano scenario might not supply us with an illustration of the mistake in Block's approach. One might think that this fact does make a difference, since the machine might seem to have the *capacity* to respond with intelligent answers, while Christian might not seem to have the capacity to answer poetically (compare what Block writes about a similar example on page 22). However, although Christian can answer poetically only because Cyrano *is there*

*beside him*, telling him what to say, Block's machine can answer intelligently only because its programmers *were there beside it*, programming it. Had Cyrano supplied Christian, or the programmers the machine, with silly answers, neither would have responded sensibly. So it seems that we should either allow both or deny both their corresponding capacities.

Be that as it may, what is important is that in both cases we do not have to look into internal processes but to linguistic behavior in order to determine whether the one is a poet and the other intelligent. What we need to realize is that Christian and Block's machine reproduce others' responses.

I conclude, then, that Block was wrong in concluding "that whether behavior is intelligent behavior is in part a matter of how it is produced." Block has thus failed to demonstrate that the nature of internal processes can be conceptually relevant to psychology. He was not justified in maintaining that what he called 'psychologism' is true.

6. Let us now return to criteria behaviorism. As was said above, Block intended not only to establish what he called 'psychologism', but also to refute behaviorism. And although, as I have argued, he failed to do the former, one might still think that he succeeded in doing the latter. However, I shall show that Block's example, rather than refuting criteria behaviorism, actually supports it.

Block's machine lacks intelligence because of its limited *capacities*. Block's machine can provide intelligent answers only in *exceptional circumstances*, when it has been programmed in a very specific way. His machine does not have any independent intellectual capacity, but has to be given the right answers in advance. Had it been provided with senseless answers (as may be the case if its programmers occasionally made mistakes), it would have responded accordingly. Block's machine does not *in general* have the required ability, and it therefore lacks an intellectual capacity. *That* is the reason why it lacks intelligence.

Thus, behaviorist considerations, regarding one's capacities—what one can do and in what circumstances—determine that Block's machine is devoid of intelligence. Consequently, Block's refutation of the neo-Turing Test is not a refutation of criteria behaviorism.

Block was wrong to claim that his machine has the capacity of an intelligent person. An intelligent person does not have to be always given the right answers in advance, while Block's machine does. Block made his mistake because, having the Turing Test in mind, he thought of his machine from its interrogator's point of view. And the interrogator cannot, indeed, distinguish between the machine and an intelligent person. But the interrogator does not have all the information required to determine the machine's capacities. To determine that, we need to know how the machine interacts with its environment in other circumstances as well—we need to know what happened to it in the past, and not only what it can do at present . And the fact that its interaction with its programmers determines its responses is relevant for determining its capacity to respond intelligently.

7. If my analysis in this paper is correct, then one influential argument against behaviorism is ineffective. But this is obviously insufficient for reinstating behaviorism, even in its criteria version. Unlike Cyrano's new-born Hercules, neo-behaviorism has not merely two serpents to overcome, but a Hydra of argued Doubts. Yet I hope a first

step in this labor has been made.

8. *Appendix*

A Hydra indeed! This paper has only been published and the creature has grown a new head in the old one's place: Ned Block, having read my paper, wrote to me in response. I have also discussed the subject with him when we met in Pécs in May 2006. Since our correspondence was private, I shall not quote from it here. Instead, I shall bring Block's response in my own words and try to reply to it. And of course, had Block written the objection himself, it may have been more powerful.

I have described a machine constructed by engineers and programmers, and explained why such a machine would be devoid of intelligence. However, couldn't an identical machine be produced by some kind of cosmic accident, say, without any directing intelligent hand? My arguments would not apply to such a machine, but it seems it has to be just as devoid of intelligence as the one I considered, and for the same reason—after all, these machines are molecule for molecule identical. So my explanation seems wrong.

But first, let us recall the externalist position, widely maintained in contemporary philosophical literature. Two creatures might be physically identical yet mentally different, due to their different histories or environments (Putnam, 1975, and subsequent literature; the accidental machine is reminiscent of Davidson's Swampman). So whatever are our verdict and argument regarding the constructed machine, those regarding the accidental one may differ. The accidental machine might not be devoid of intelligence, or, if it is, that might be due to a reason different from the one due to which the constructed machine is. Consequently, if, ignoring the accidental machine's case, no fault is found in my arguments against Block's claim with respect to his constructed machine, these arguments may still stand. And in our correspondence and discussion Block did not note any such fault, nor am I aware of any. I therefore still think that I have shown that the constructed machine is devoid of intelligence not in virtue of its internal processing but because it responds only with answers given to it in advance.

Now, what about the accidental machine? Well, in fact I think *it is* intelligent. Think about the matter this way: we witness an explosion, out of which emerges a machine that answers intelligently all our questions: wouldn't we consider it intelligent? Would we think that we should first inquire into its internal processes? The latter doesn't seem reasonable to me.

Of course, Block could (and did) respond that I am begging the question: I am just maintaining what seems to me right, and not supplying any argument. I have not shown that in this case we needn't take internal processes into consideration. But then, this response equally applies to him: if he were to claim that the machine is devoid of intelligence by virtue of its internal processes, he would just be maintaining what seems right to him: neither did *he* show that in this case we *should* take internal processes into consideration. Moreover, his thought experiments were supposed to rely on our intuitions, that is, on our understanding of how our concepts apply in the situations described; and on the basis of these intuitions we should have *inferred* whether internal processes are relevant to intelligence. So I am not sure either of us *need* supply any argument, in the sense mentioned, for or against the relevance of internal processes.

Yet more can be said in support of my description of the cosmic-accident

machine. First, by contrast to the constructed machine, all the responses this machine supplies are *new ones*: it does not reproduce responses given to it by anyone else. What more is required for intelligence? Secondly, suppose we do check its internal processes, and then discover that by some extraordinary chance the machine is such that it has in advance all responses to all possible questions, where these responses take into consideration earlier questions and replies: wouldn't that be, in fact, a good reason for thinking the machine *is* intelligent? So I am not convinced that Block's processes examination supports his claim in this case. (Note that on my view the structure's examination would convince us the machine is intelligent not because the structure is *conceptually* relevant, but because we can infer from it how the machine would *react* in all possible circumstances.)

Of course, if either the perfect constructed machine or the perfect accidental one were really possible, and not just logical possibilities, then our ordinary concepts of intelligence and mentality might not be quite useful to describe the situations in which they existed. But since neither machine is, this should not concern us here.

My judgment about the accidental machine thus fits my judgment concerning the constructed one. We judged the latter to be devoid of intelligence because someone else supplied all its replies. This is not the case with the former, and we indeed judge it to be intelligent.

So by contrast to Block, I think that our concepts do not supply the same verdict on the accidental machine as they did on the constructed one. And again—and most importantly, given the purpose of this paper—no mention need be made of the accidental machine's internal processes in order to explain this applicability. No case has therefore been made for Block's 'psychologism'.

**Acknowledgements**

I am indebted to Peter Hacker, John Hyman, Roger Teichmann, Ruth Weintraub and two anonymous referees for comments on earlier versions of this paper.

**References**

Block, N. (1980). What intuitions about homunculi don't show. *Behavioral and Brain Sciences*, *3*, 425-426.

Block, N. (1981). Psychologism and behaviorism. *Philosophical Review*, *90*, 5-43.

Braddon-Mitchell, D. & Jackson, F. (1996). *The philosophy of mind and cognition*. Oxford: Blackwell.

Putnam, H. (1975) The Meaning of 'Meaning'. Reprinted in his *Mind, Language and Reality*. Cambridge: Cambridge University Press, 215-71.

Putnam, H. (1975). *Mind, language and reality*. Cambridge: Cambridge University Press.

Rostand, E. (1951). *Cyrano de Bergerac*, (B. Hooker, Trans.). New York: Random House.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417-424.

Wittgenstein, L. (1958). *Philosophical investigations* (G.E.M. Anscombe, Trans.).

Oxford: Basil Blackwell.
Wittgenstein, L. (1967). *Zettel* (G.E.M. Anscombe, Trans.). Berkeley and Los Angeles: University of California Press.