CDD: 153

INTELLIGENCE, BEHAVIOR AND INTERNAL PROCESSING

J.I. BIRO
The University of Florida, U.S.A.

Em trabalhos recentes Ned Block fornece alguns argumentos novos para mostrar que "o psicologismo é verdadeiro e portanto uma análise behaviorista da inteligência que seja incompatível com este último tem de ser falsa". Ele elabora um experimento mental no qual uma máquina é programada para exibir comportamento aparentemente inteligente e mostra em que sentido nossa intuição leva-nos a concluir que tal máquina não é realmente inteligente. Tal intuição origina-se no fato de que a máquina é concebida como operando com processos internos aos quais falta uma estrutura interior dotada de um certo tipo de complexidade e que tais processos internos são mecânicos, refletindo apenas a inteligência do programador.

A argumentação contra Block é feita em duas direções: em primeiro lugar mostra-se que a origem e a natureza de nossa intuição é incorreta; em segundo que as conclusões que ele extrai a partir do psicologismo não procedem. Embora a intuição que Block invoca possa mostrar que o behaviorismo é falso, isso não equivale a provar que o psicologismo no sentido por ele pretendido seja verdadeiro.

Ned Block has recently adduced some new arguments to show that "psychologism is true and thus a natural behaviorist analysis of intelligence that is incompatible with psychologism is false". He introduces a thought experiment in which a machine is programmed to exhibit intelligent-seeming behavior and appeals to our intuition that such a machine is nevertheless not really intelligent; he traces that intuition to the fact that the machine is being thought of as operating with internal processes that, first, lack a certain internal structure with a certain kind of complexity and, second, are in a certain sense mechanical, reflecting only the programmer's intelligence.

I argue against Block both that he mistakes the source and nature of our intuition and that the conclusion he draws concerning psychologism does not follow. While the intuition he appeals to may show that behaviorism is false, that is not equivalent to showing that psychologism in the sense intended by Block is true.

> "They felt that if you could program something, then the machine was not 'really' doing it ..."
> (Marvin Minsky, describing reactions to early AI programs in *The New Yorker*, Dec. 14, 1981)

I

Does being intelligent consist simply in having the ability to behave in certain ways, to exhibit a pattern in one's behavior that we would regard as the normal manifestation of intelligence? Or does it require that the behavior in question be the product of certain sorts of internal processes? Opting for the second alternative amounts to espousing *psychologism*, in the broad, general sense in which Ned Block opposes that doctrine to *behaviorism* (Block, 1981). Block argues that "... psychologism is true and thus a natural behavioralist analysis of intelligence that is incompatible with psychologism is false." (Block 1981, p. 5) While he cautions that the label 'psychologism', as he uses it, should not be taken to denote any very specific or very rich doctrine, it is clear enough that even in this general sense, it is supposed to be incompatible with behaviorism of any sort. If he is right about the connection between genuine intelligence and the production of a system's behavior by internal processes with certain properties, then no account focusing just on properties of that behavior, on features merely of the output of internal processes, can be adequate.

Block bases his verdict in favour of psychologism on a thought experiment, claiming that our intuitive reaction to it is sufficient to show behaviorism to be false and psychologism to be true. Both the thought experiment and the use Block makes of it have much in common with John Searle's well-known "Chinese-room" argument, and, of course, the conclusion drawn is much the same (Searle, 1980). Although Block's immediate target is behaviorism and, more specifically, what he calls

the "Turing test conception of intelligence", whereas Searle's argument is aimed at functionalist theories of intentionality generally, it is clear that both the substantive issues and the method of approaching them are the same. I choose to discuss Block's version of the argument because it is formulated in a way that will allow me to look at what may lie behind the intuition both he and Searle rely on, rather than just taking it at face value. This, in turn, will make it possible to raise some questions about the reliability of that intuition as a basis for an argument for psychologism.

Block claims that his argument decisively refutes natural behavioristic accounts of intelligence better than the standard arguments against behaviorism, no matter how these accounts are amended and elaborated[1]. His own argument involves describing a machine constructed so as to be capable of exhibiting behavior of a kind we would describe as intelligent if found in a human being. When we do so characterize human behavior, we take it that the exhibiting of such behavior entails the possession of the capacity to so behave and that that, in turn, entails the possession of intelligence. Block's strategy is to challenge the second stage of this inference, that from (even) having the capacity to behave in an (apparently) intelligent way to (really) being intelligent.

The capacity to behave intelligently must be distinguished from the mere exhibition of intelligent behavior, since the occurrence of the latter on an occasion may be accidental. Only if tests designed to elicit intelligent behavior are consistently passed do we have strong

[1]Block lists three standard objections to behaviorism: "the Chisholm-Geach objection" (that the conditions under which a particular mental state will lead to a particular behavioral disposition must include other mental states), "the perfect actor objection" (that no behavioral disposition is necessary or sufficient for at least some states, such as pain, given the possibility of a community of beings who can pretend), and the "brain-in-a-vat objection" (that beings unable to behave in the usual sense of 'behave' can be in states such as pain). Since my criticisms of Block do not depend on his treatment of these objections, I shall say no more about them here.

evidence that the entity in question possesses the capacity for intelligent behavior, rather than just fortuitously behaving on some occasion in the way a system with that capacity would behave on an occasion of that sort. But even having the capacity to behave intelligently is insufficient for actual intelligence in the behaver. That, in a nutshell, is the burden of Block's argument. It is possible, he claims to be able to show us, for a system to be capable of exhibiting intelligent behavior and for that system not be genuinely intelligent. He describes a system – a conversational machine – which he takes to illustrate this. It is a machine in which internal processes go on of a sort sufficient not just for the manifestation of behavior naturally described as intelligent, but also for that behavior to be the manifestation of a capacity to produce such behavior. Yet, Block claims, the machine is clearly such that we would intuitively deny that it is really intelligent.

The machine Block ask us to imagine is programmed in the following way. For every sentence an interrogator (say, one conducting a Turing test) produces, the machine produces a sentence in response, one that constitutes a natural and intelligible continuation of the "conversation". It does so by looking up the input sentence on a string of sentences it has been provided with by its programmer, a string in which every sentence is succeeded by a natural conversationally acceptable response: it then returns that successor sentence as its output. The machine will, of course, need many such strings, each differing from the others, to reflect the different possible sensible conversations, in order to have enough potential variety in its responses not to be found out by a suspicious interrogator. Each time the interrogator (better, interlocutor, since not all input sentences need be questions) takes his turn, the machine searches among its strings and selects one among those whose sequence of sentences matches the "conversation" up to that time. It then selects one of these strings and makes the next move, conducting a similar search prior to each succeeding move. Given enough strings, the

machine can pass a Turing test of any finite length[2]. Yet, Block claims, it is obvious that the string-searcher is stupid; he thus concludes "that the capacity to emit sensible responses is not sufficient for intelligence" (1991, p. 21).

In this paper I want to examine Block's argument and ask a number of questions about it. First, allowing that our intuition is as he claims, what exactly does that show about behaviorism and psychologism, respectively? (The answer will help clarify the relation between these two doctrines, a relation not as simple as Block evidently takes it to be.) Second, what, if anything, is it about the machine Block describes that precludes it from having real intelligence? We will find that the answer to this question is not entirely clear and that of the various answers suggested by Block's discussion – some having to do with the character of the machine's internal processes, some with the origin of the program governing them – none is strong enough to support the claim Block takes the intuition to establish, namely, that such a machine *cannot* be intelligent as long as it lacks internal processes of a certain sort. Only reliance on what one may call the "causal isomorphism principle," can yield such a conclusion. But that principle already embodies the psychologism the argument it would be part of was designed to establish. Hence, it would both beg the question and make Block's appeal to his thought experiment redundant.

[2]Block also describes a tree-searcher version of the machine, where the strings are relaced by branches, the machine making a new move by producing just the sentence appearing next after the *interrogator's* choice of a branch at each node (1991, p. 20). The two machines seem rather different, actually, in that the string-searcher appears to have a "choice" of moves the tree-searcher does not. For reasons that will be apparent later, the string-searcher is for this reason more likely to give the appearance of a certain creativity we associate with intelligent humans than the tree-searcher. Still, Block would of course say that, unlike with humans, it *is* only an appearance.

II

Suppose, for the sake of argument at least, that Block is right that in our intuitive judgement a machine of the sort he describes fails to meet our standard for the possession of genuine intelligence. (I shall discuss later the sources and status of the intuition on which that judgement is based and, in particular, its ability to support such a judgement.) We may still be making this judgement for one of two very different reasons. One would be that we think that the properties attributed to the machine are insufficient for the possession of intelligence. The other would be that we think that those features are incompatible with the possession of intelligence. These different reasons in fact make for two different judgements. Both are in conflict with a "behaviorist analysis" of intelligence. But only the second is relevant to establishing psychologism in any interesting sense. This is because the first leaves open the possibility that intelligence is *in fact* a property grounded in the very features we have attributed to our machine, even if it is not a conceptual truth that it is. The second, by contrast, declares that it is a conceptual truth that such cannot be the case. This amounts to claiming that there are some other properties that any system must have if it is to be intelligent. If, and only if, these properties include some involving the nature of the internal mechanism that causally underlies the behavioral output of the system, do we have an argument for psychologism. (We will see later that it is doubtful that the intuition Block is relying on can sustain such a strong claim.)

Block's discussion does not respect the difference just noted. He appears to think that to establish that analytic behaviorism is false is, *ipso facto*, to establish that psychologism is true. But even if his thought experiment establishes the first, it cannot, by itself, establish the second.

What else is needed to have an argument of the sort Block wants *for* psychologism? We could adopt a principle according to which the

causal relations we naively suppose to obtain between the mental states we take to underlie behavioral capacities – among them, conversational ones – must have counterparts in the physical mechanisms the workings of which give rise to those capacities. Such a principle – I have suggested calling it the "causal isomorphism principle" (CIP) – would block the possibility of interpreting our intuitive judgement in the first of the two ways just outlined, an interpretation on which, while it arguably refutes analytic behaviorism, fails to establish psychologism. There is, admittedly, something plausible about a principle such as CIP. It is not unnatural to assume that whatever causal structure is conceptually part of our thinking about intelligent behavior must be mirrored in whatever mechanism produces that behavior. However, it does not take much reflection to see that the principle cannot be a necessary truth, for we cannot rule out the possibility that instead of the naively posited mental causal structure what we have at the level of the underlying physical mechanism is some other causal structure nomologically sufficient to produce the behavior in question, even to constitute the behavioral capacity that may give rise to the behavior.

There may, of course, be good reasons for positing CIP (or something like it) as an empirical hypothesis about how certain kinds of behavioral capacities are in fact grounded. The addition of CIP (and *only* that) can turn our verdict on Block's machine into the second kind of judgement I distinguished earlier, one that is strong enough to establish the truth of psychologism. But then there is really no need for the thought experiment at all, and no use for it in any *argument* for psychologism. CIP already embodies psychologism, and the truth of that doctrine would ride entirely on the strength of the arguments for CIP.

Could we, however, think of the thought experiment as itself providing at least part of that argument? (Obviously, if it provided all of it, it would not need the addition of CIP as an independent principle

to sustain the strong psychologistic conclusion. But we have seen that it does.) To see what, if anything, we can get out of the intuitive judgement Block claims we would make about his machine that might tend to support CIP, we must now turn to a closer examination of what Block's discussion suggests about the sources of that judgement.

III

I shall begin with a more detailed description of Block's thought experiment than I have given so far. Limiting the discussion for the most part to conversational intelligence manifested in responses to verbal stimuli, Block considers what he calls the "neo-Turing Test conception of intelligence"[3]. "Intelligence", writes Block, "is the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be" (1981, p. 11). Block begins his argument by describing very simple systems, such as Weizenbaum's ELIZA, about which there will be a consensus of intuitions: they are not intelligent (Weizenbaum, 1965). These simple systems are capable of producing sensible-seeming responses to a fairly wide range of questions on a given topic, especially if programmed by a few trick responses such as replying to a question by asking another question[4]. Block then ask us to consider "some more complex (but nonetheless unintelligent)" system. If even the former can sometimes fool gullible people, as they apparently can, how much more likely that the latter might! Indeed, if we make them complex enough, they might fool us all, no matter how cautious and sceptical we are (Block 1981, p. 11).

[3]This limitation, Block claims, is inessential, and is adopted only for simplicity's sake. The argument is alleged to be generalizable to intelligent behavior of any sort (1991 p. 23). I shall touch on this question in the last section of this paper.

[4]Equipped with a fairly simple bag of tricks, ELIZA can play a plausible psychiatrist. If you mention your father, it will come back with "What else comes to mind when you think of your father?"; if you mention your boyfriend, it may later respond to just about anything with "Does that have anything to do with the fact that your boyfriend made you come here?", if all else fails, it may demand "Who is the psychiatrist here, you or me?" (1965, p. 10).

Now imagine a system much more complex than ELIZA, though relying on similarly mechanical means (and perhaps similar tricks) to produce its sensible-seeming responses. Such a system may even pass a Turing test of considerable length. Yet, would we not say that it, too, is unintelligent? If we would, a Turing test conception of intelligence cannot be an adequate one. Something more than just passing the test must be required for the proper attribution of intelligence. Something besides the ability to "fool most any human judge" (Block 1981, p. 10).

But exactly why is this so? The first thing to notice is that it is not entirely clear on just what the intuition Block is appealing to, namely, that systems of a certain kind (of whatever degree of complexity) are obviously unintelligent, is based. There are at least two different elements in his description of such systems that we must separate, and separate more sharply than he sometimes does, before we can form a clear estimate of the source and reliability of this intuition. First, the responses of the systems are pre-determined, already waiting there cut-and-dried, as it were, before the stimulus occurs. Thus they lack a certain *spontaneity* (and would be, in the absence of a randomizer, completely predictable in principle.) Second, these pre-fabricated responses are placed there by some programmer outside, and distinct from, the system itself.

Block takes some pains, as I have said, to keep these two features distinguished (e.g. 1981 p. 27, fn. 21). But in describing his examples, his language perhaps inevitably echoes both. It is therefore hard to tell which feature is playing what role in the way we intuitively respond to the cases he describes. Take for example his suggestion that some more complex version of ELIZA may be able to fool all human judges no matter how careful and sceptical. This sounds plausible, yet presumably it would not be the system that was fooling human judges in such a case but, at best, its designers. It does not make sense to suppose that a system could succeed in really fooling us (as opposed to just behaving in a way that misleads us), without really being intelligent. That the

ability to fool people presupposes intelligence is not something Block would deny, of course. Indeed, he would insist precisely that in such a case it would be the designers who would be fooling us, rather than the system. So talk of the latter fooling us is just not to be taken literally, any more than would be saying that the bright sunshine I see through my window fools me into thinking that it is a warm day, that the chameleon's change of color or the ordinary butterfly's Batesian mimicry of the monarch fools their respective predators.

Having described his conversational machine – the string - or tree-searcher – Block in fact concludes:

> 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* (1981, p. 21, Block's emphasis)

Again:

> The trouble with the neo-Turing Test conception of intelligence ... is precisely that it does not allow us to distinguish between behavior that reflects a machine's *own* intelligence, and behavior that reflects *only the intelligence of the machine's programmers* (1981 p. 25, Block's emphases).

How much should we be impressed by the second-hand character of the intelligence, emphasized so strongly by Block? People (at least some of them) are intelligent, yet we could say that all the intelligence *they* exhibit (such as it is) is that of their "programmer". Something like this would have been said by many philosophers until relatively recently and is still believed by many non-philosophers. The only reason why it now seems odd to say it is that the traditional notion of God, a sort of super-programmer, is no longer regarded as a viable explanatory notion. But in fact we have something to replace it with, namely evolutionary

theory, which tries to explain the mechanism by which intelligence is "programmed" into us. It does so, of course, without postulating a purposeful programmer with his own ends distinct from the ends of the entities it programs in ways designed to achieve these ends. In both these ways, it is different from divine-creation theories. Such theories (no less than those that appeal to an intelligent homunculus *inside* the systems as the source of the intelligence of the whole), embody what has been called the exempt-agent fallacy: they locate the source of the creature's intelligence in some other creature, inside or outside[5]. With evolutionary theory, we can have our "programmer" and avoid making at least part of this mistake: evolution is blind.

Block's claim that even if his string-searcher has the capacity to exhibit behavior we would, in a human being, describe, as intelligent it is still not really intelligent certainly seems plausible at first sight. I suggest that this plausibility comes in part from the fact that we think of its programmer as intelligent in a certain way, namely, as having goals and purposes of his own, and we think of him as having a kind of understanding which goes with the possession of such. Should we, though, suppose that in thinking of another creature, this time of a human one, we must also postulate some analog of a programmer intelligent in the special sense described above, one having its own goals purposes? It would then follow that since in the human case we obviously cannot place such a programmer outside the creature, we must place him inside: the homunculus (non-)solution. But there is an alternative. We can place a "programmer" – evolution – not intelligent in any sense, outside the human creature, and in so doing, free ourselves to regard the intelligence of the creature as neither exhausted by intelligent-looking behavior, nor borrowed from some other "really" intelligent creature, inside or outside. This is precisely the role evolution can be seen as playing *vis-à-vis* human intelligence.

[5]See D.C. Dennett (1977), esp. pp. 101-2, Dennett, (1978a) esp. pp. 123-5, and J.I. Biro, (1981).

What we need to avoid is what might be called the *banking* model of intelligence, one on which one party (the programmer) lends from its intelligence capital to the other (the machine or human being). The model presupposes that any adequate (causal?) explanation of the possession by some entity of some characteristic must proceed by way of locating that characteristic in some other entity and by then telling a story about how the latter imparts the characteristic in question to the former. No matter how complicated the story gets, the assumption from which the argument starts is always the same: there can be no emergent characteristics[6]. The alternative explanation in terms of the interplay of some set of characteristics giving rise to the appearance of a new one is not considered[7].

Homunculus theories embody the banking model in a special way: the banker is thought of as being inside its client, the creature whose intelligence we are explaining. The creature *behaves* intelligently, all right, but we think that is not enough to justify saying that it possesses the characteristic of intelligence. We are then driven to think of some "part" of the creature as being the lender – of last resort – in the case; we think that that component of the creature must be such that *it* does possess the characteristic in question (is *really* intelligent), otherwise how could it lend it to the creature?

---

[6]The "banking model" is ubiquitous, present already in Descartes' Third Meditation. An important modern view threatened by it is Fodor's. (See, for example, J. Fodor (1978), pp. 228-47, and his (1980) pp. 63-109, esp. p. 65). The model fails to fit intelligence only in that financial capital, unlike intelligence "capital" is normally conceived of as finite, whereas the programmer's intelligence (whether human or divine) is not thought of as diminished by the loans he makes. But note how well other aspects of the analogy hold: intelligence is lent or given *for* a purpose (of the borrower's) and *on* a purpose (of the lender's), and is given with strings attached in terms of certain expectations of performance, on pain of arbitrary intervention by the lender if these expectations (tailored to fit *his* purposes) are frustrated. The model, of course, applies to much more than just intelligence: it is the traditional and still powerful model for understanding life.

[7]In speaking of 'appearance' here I do not mean that the characteristic is a merely apparent one: it is as real as any other, including those others which give rise to it. I mean 'advent'.

I am not saying that the banking model is inappropriate for Block's machine. On the contrary, it is clearly the right way to describe *it*. My point is rather that it is not the (only) one to be given for systems whose internal processing is like that of the string-searcher. So, if *people* turned out to be like that, we would have more options than denying them genuine intelligence, as Block asks us to deny to the string-searcher. Before Darwin, it seemed no more plausible to people that we – our very life, let alone our intelligence – could be the result of blind process without a "banker" lurking in the background, than it now does that a machine like Block's might. They were wrong[8].

Now how does all this affect Block's main thesis, that psychologism is true? Clearly enough, psychologism may *be* true, for all I have said. Indeed, we may even concede to Block that his argument based on the machine analogy does show that the mere exhibition of intelligent behavior, even the possession of the *capacity* to so behave, does not entail possession of intelligence in some stronger sense (perhaps equivalent to what used to be called a faculty). But the argument shows this in a very misleading way. It suggests that what is missing is a certain etiology of the behavior in which something other than the behaver (though perhaps something that is part of him) endowed with (the faculty of) intelligence plays a crucial causal role. In doing so, it may raise all the old problems of regress which behaviorism was in part an attempt to by-pass.

[8] Could *we* be similarly wrong about the possibility of the evolution of machine-life? No, because we know something about the materials machines are made of, and everything we know tells us that the properties of those materials are not capable of giving rise to the kinds of properties (vegetative, reproductive, locomotive and, maybe, cognitive) we associate with life. We also have some theories about the connections between the various kinds of properties we can attribute to things from each of the three explanatory "stances" distinguished by Dennett (see Dennett 1971). We know that some physical properties cannot realize some design properties and that some designs are not suitable for realizing some intentional properties. (Imagine trying to build a mini-computer from wood.) So the point is not (of course) that we might come to think of Block's string-searcher as being the product of evolution. It is that something with genuine intelligence and internal processes possibly just like the searcher – a human being – *is*.

Block does attempt to insulate what he claims is his main argument from this misleading feature of his example. He suggests that the fact his string-searching machine is imagined as having been designed by intelligent programmers is incidental, and that we could think of it as the product of a cosmic accident (1981 p. 27, fn 21). In any case, he allows that even a machine designed by intelligent beings might be thought of as having intelligence *of its own* – rather than just by courtesy of its creators – as long as its internal processes were *different* from those of the string-searcher. So let us consider whether a removal of the second-hand aspect of Block's example leaves enough for his argument in favor of psychologism.

Thus one could insist that the aspects of Block's description of his thought experiment that I have been suggesting are responsible for the intuition we (are supposed to) have about it – the second-hand character of the string-searcher's responses, and their lack of spontaneity – are really incidental. We could – so this objection goes – re-state the thought experiment in a way that did not involve these features at all: as we have just seen Block suggest, we could think of the string-searcher as coming into being through a cosmic accident[9]. All that is necessary to the thought experiment – and thus to Block's argument – is that the searcher produces its intelligent-seeming responses without *understanding* them (or, of course, the questions to which they are responses): our intuition would still be that it is not *really* intelligent[10].

I think, however, that this objection mis-locates the point at which the alleged intuition has to be evaluated. Indeed, if we already know that the searcher lacks understanding, we will judge it to be without intelligence. But that is really just saying the same thing twice, in dif-

[9]It is actually not clear that this would remove the second feature. Even if not programmed to operate the way Block describes, as long as the searcher did so operate, it might still seem to lack spontaneity.

[10]This is essentially Searle's version of the argument. As I said in section I, I chose to discuss Block's version precisely because in it we can see beyond the bare intuition to the connection between it and some other, challengeable, assumptions about what understanding itself requires.

ferent words. The intuition Block must ask us to share is that a system whose internal processes have the features the string-searcher's have *cannot* be said to *understand*. This judgement, unlike the "inference" from lack of understanding to lack of intelligence, is far from being a trivial one. But, as I argued in section I, one would need to be shown that processing done in the way the string-searcher does it is *incompatible* with having understanding, and, hence, intelligence. Nothing Block says shows this.

Why should we believe that a system with internal processes like the string-searcher's could not understand what it was doing in the same sense in which we understand what we do? Do we know that our internal processess are not like its? And even if we did, would that show that doing it that way was any less compatible with understanding than doing it the way we do, or some other? Thus it seems that we cannot side-step the question, just what is it about Block's description of the string-searcher's way of doing things that might make one think that it must be dumb? I can see nothing other than the features the line of objection I am discussing suggests are merely incidental features of that description. I have argued that they are not so, but that when examined, they fail to justify the intuition Block thinks natural and appropriate.

IV

Let us look more closely at the thought experiment with its second-hand aspect removed, to see whether it can still support Block's argument, as he suggests. We still have a description of internal processes which are indubitably mechanical. Is it just as clear now, without further argument, that this machine, with its enormously more complicated output and its correspondingly more complex internal processing, will evoke the same intuition as did relatively simple programs like ELIZA? Or perhaps our willingness to regard a system as intelligent does depend on the complexity of that system; perhaps as we think of

more and more complex systems there is a point beyond which we no longer want to characterize the system as *un*intelligent.

Block's point is precisely that there is no such point: that however complex a system we envisage, it will still be *obviously* unintelligent. If it is not the second-hand feature that backs up this claim, what could? Here Block's answer is not easy to find. It seems that his intuition that a system whose processing is of the sort he describes (no matter how complex) would be obviously unintelligent, is prompted by a certain lack of *spontaneity* or *creativity*. (While this does not necessarily lead to predictability – remember the possibility of a randomizer – it does naturally carry that suggestion, and in so doing muddies the waters even further.) This lack of spontaneity is to be distinguished from the lack of *autonomy* that would result from the responses being second-hand, placed in the system by a programmer; we are now thinking of the searcher as having come to be (the way it is) naturally, perhaps by accident. It may still be argued that the mechanical and predictable character of its responses prevents us from attributing intelligence to it. This seems to be Block's intuition and he expects that everyone will share it. But, again, I am not so sure that I do. In section V, I shall argue that this intuition is perhaps less reliable than Block thinks and that it can be shaken by consideration of autonomous systems which are unarguably mechanical and yet not obviously different from us, systems which if we did not know *how* they worked – as we do not with us – could easily give the same *appearance* of spontaneity and creativity that we give.

Before describing such a system, however, I would like to take up briefly the one argument Block discusses that seems to buttress the intuition that mechanical systems cannot be intelligent. This is the argument from the danger of combinatorial explosion, and Block's discussion of it in connection with the "*amended* neo-Turing Test" (defended by Dennett, for example) comes as close to giving an argument, rather

than relying on bare intuition, as anything in his paper. The amended neo-Turing Test runs as follows:

> Intelligence is the capacity to emit sensible sequences of responses to stimuli, *so long as this is accomplished in a way that averts exponential explosion of search* (1981 p. 38, Block's emphasis).

Notice that the amended test says nothing about *how* combinatorial explosion of search is to be averted. Even so, Block suggests that in allowing even this much constraint on the internal processing of the system, the test concedes the argument to the defender of psychologism:

> The *amended* neo-Turing Test conception ... [places] a *condition on the internal processing* (albeit a minimal one), viz., that it not be explosive (1981 p. 39, Block's emphasis).

So the amended neo-Turing Test conception *does* characterize intelligence partly with respect to its internal etiology; hence, the amended neo-Turing Test conception is psychologistic (Block, 1981 p. 49).

We need to ask, however, whether "characterizing intelligence partly with respect to its internal etiology" is psychologistic in a sufficiently interesting sense. Apart from the vagueness of the phrase 'partly with respect to', all the amended neo-Turing Test concedes is that there must be *some* means of averting combinatorial explosion. It makes no commitment to any particular way of doing so over any other and, most importantly, no commitment to any interestingly *non-mechanical* way. Here Block may say that the string-searcher's way of averting explosion could only be *postponing* rather than *avoiding* it (1981 p. 39). It may postpone it long enough for the machine in question to pass a Turing Test of a reasonable length, but that is not the same thing as avoiding it altogether, in the sense of using methods and resources which would *never* lead to such explosion. If we bear in mind this ambiguity in the term 'averts' which the amended test uses, we are faced with a

dilemma. If we are satisfied with postponement, a machine may count as intelligent. If we insist on genuine avoidance, *we* may not. This is because there is no conclusive way of showing that the methods *we* use do not merely postpone, rather than avoid, explosion, since we never face tasks where the difference would matter or could show up. (We could never be given a Turing test of sufficient length, so that while we may *actually* use methods that would lead to combinatorial explosion if certain demands were put on them, they can handle the actual demands put on them without such explosion.)

Block regards this as a dilemma for the amended neo-Turing Test. But why can its defender not reply as follows? We do not know – and perhaps can never find out – whether human beings use methods which merely postpone combinatorial explosion, rather than ones which altogether avoid it (if such methods there be). So, we should be prepared to say of a machine that passes a neo-Turing Test of reasonable length that it *is* intelligent for precisely the same reason that we are prepared to say that human beings are intelligent. There is no reason to demand *more* of machines and other non-human systems than we demand of people, indeed – bearing animals in mind – perhaps we should not even demand as much. If we are prepared to accord intelligence to a person in the face of our uncertainty about how he manages to avoid failing the amended neo-Turing Test (whether he does so by postponement or by avoidance), we have no reason to describe a machine as *un*intelligent merely because we know that *it* manages to avoid failing the same test by mere postponement[11]. For should it turn out that people also

[11] It may be objected that the cases *are* different in that we at least know that the machine couldn't be really avoiding, whereas human beings *might* be. But, first, Block himself seems to concede that it is not clear that the string-searcher with its mechanical methods couldn't be regarded as avoiding explosion altogether, in that it is nomologically possible (in a universe with infinitely divisible matter) that its memory be *indefinitely* large (1981 pp. 31-2). With such a memory, in fact, the distinction between avoiding and merely postponing collapses. Second, even if we knew that the string-searcher merely postpones, attributing intelligence to humans and not to such machines on the grounds that the former *might* not be seems rather lame.

merely postpone, we would undoubtedly continue to describe *them* as intelligent. We *know* they are. But if it would make no difference to our description of human beings, why should it – how can it – make a difference to our description of machines?

So it seems that while it is true that the amended neo-Turing Test does place a condition on the internal processing of a system, that condition is not strong enough to constitute a concession of psychologism in any interesting sense. The condition cannot require that the processing be such that it should never lead to explosion, as opposed to requiring only that it merely postpone it long enough, for we cannot say with confidence that we ourselves do the first, rather than the second. But if it is possible that mere indefinite postponing will do, it would have to be shown that there can be no purely mechanical way of doing *that*. (In fact, it is likely that there is, especially if loops are allowed.) In either case, we cannot conclude that any machine or human being who satisfies the amended test *must* do so by relying on internal processes different in kind from those mechanical ones Block's conversational machine relies on. But then the amended neo-Turing Test does provide a natural behaviorist analysis of conversational intelligence in spite of the fact it puts a constraint on internal processing. Passing the amended test involves doing so without intolerable explosion of search. Obviously, if explosion were actually to occur, the test would be failed: no sensible-seeming responses would be forthcoming. Anything that actually passes, then, has a title to intelligence. We cannot infer from this anything that passes cannot be employing internal processes of the sort Block claims are obviously stupid; hence we can infer no interesting psychologism.

V

Let us now briefly consider a system that does what it does in a clearly mechanical way and yet may seem to have, just as we do, not

only the autonomy but also the spontaneity and creativity we think a genuinely intelligent system should.

To begin with, consider an actual system of inputs the connections among which are based on clearly mechanical internal processes. Such a system is Rubik's notorious and frightening cube. Given that the cube is in a certain state, let us say S: a certain I_k (provided by the player) will move into one of the several other states, $S_1 \ldots S_n$. There are, in fact, 27 possible states into which a 54-square cube can directly go from any given state, *and this for mechanical reasons*, because of how its insides are. Thus a cube in state S; receiving input I_k will go into one of the states in the set $\{S_1, \ldots, S_{27}\}$ depending on the character of the input (i.e., which way it is twisted by the player).

So far, our cube is purely passive, deterministic and thus certainly not autonomous. What should be noticed, however, is that there is a limited number of possible output states or responses on each occasion of input and that that number is far smaller than the possible states of the cube ($4 \text{ X } 10^{19}$ for a 54-square one), and that this limitation on the possible output states for an input is a function of the internal mechanics of the cube.

Now let us imagine a slightly different cube, one that responds to verbal inputs. Its internal processing involves decoding the verbal stimulus into a mechanical one and responding to that. There is a real sense, I think, in which a cube would immediately seem to the uninitiated to be "intelligent" and autonomous. But, of course, we are more sceptical, even if we do not know *exactly how* it is doing what it is doing. We know that there must be *some* story to be told such that its apparently autonomous and intelligent behavior will be seen for what it is: mere appearance.

Next, imagine a cube coming into being by accident (remember, Block countenances such as a possibility for his string-searcher) or evolving naturally: a living Rubik's cube. Perhaps it has a decoder to convert verbal signals into electrical impulses, so that whenever a

verbal stimulus puts it into an excited state, it "wants" or "needs" to move. The signal may not itself contain any indication as to how it should move: we are, for the moment, imagining that it is simply sound-sensitive within a certain range and that any of the verbal stimuli that excite it are equivalent. They all amount to the equivalent of a simple direction: 'Move!'. (Or we can think of it as "understanding" the signal as 'Your move!' Nothing will hang on which way we think of it.) Such a "living" cube might have the appearance of not only being autonomous, but of being spontaneous and creative in its responses as well. We could not predict which of the mechanically possible proximate states it would go into, given a certain input. It is important here to see that the decoding of the verbal stimulus involves only transforming it into, say, an electrical impulse sufficient for moving the cube into *one* of its possible proximate states. *Which* one it would move into could not be predicted from any property of the verbal input. The cube's "selection" of a move may be regarded as random, or as influenced by physical properties of it and its environment. Remember, Block's string-searcher was also said to *select* one of the many sensible strings it has in storage, but on no interesting principle (see fn. 3 above).

Now *suppose* that human beings in fact selected their responses in this quasi-random way (i.e. they *happen* to pick on one in a set of equally appropriate possibilities). Would it not look as if their responses were both sensible *and* spontaneous? Sensible, because the selection was from the set of possible sensible responses, spontaneous, because which response it was seems to be – and is – underdetermined by the verbal input. With a sufficiently complicated system, even if it operated as Block's machine does, we would not only be unable to predict its next move, but it would *seem* that there were no constraints at all on what its next state could be. But there would be, and they would be mechanical ones. It would *seem* that there are no constraints on such a system's next state for exactly the reason that it seems there are no constraints on the next conversational move of an intelligent hu-

man being: the thing is capable of *surprising*, unthought-of (by us) moves.

I am not claiming that as a matter of fact there are such mechanical contraints on a human being's next conversational move. I do not know what, if any, constraints there are, though I believe there are some and I believe that they are physical. (This leaves their precise character – electrical, chemical, or what have you – open.) Block does not know whether there are any, either. Nor am I saying that there are no non-mechanical semantic constraints on human conversational responses. But if I am right about my "living" cube, there could be a system that operated solely under mechanical constraints analogous to those placed by Block on his string-searcher and which nevertheless gave the impression of having all the autonomy, spontaneity and creativity we give the impression of having. We would be as likely to regard such a system as intelligent as we in fact are to so regard ourselves, which means that we *would* so regard it. If, were we to discover its mechanical insides, we changed our minds about it, we should be prepared to do so about ourselves, given a similar discovery. Or – the more sensible alternative – we could go on cheerfully regarding both it and ourselves as intelligent, in the face of such a discovery. What we should not do is to insist *now* that *we* should be regarded as intelligent no matter what and that *therefore* such a discovery about us is impossible.

Obviously, the analogy I have suggested with Rubik's cube goes only so far. The chief difference is that the states of the one I have asked you to imagine still have no rich semantic content, and it need take no notice of the meaning of its inputs, either, as long as these have the causal consequences I described. (Alternatively, one could say that the inputs *do* have meaning, as suggested above.) In this respect, it is, perhaps, more like an animal than a person, and it is obvious that nothing like a Turing test could be devised to test its intelligence (or what, for Block, is merely its capacity for intelligent-seeming behavior). But this is again very much like our situation *vis-à-vis* our pets. We

sometimes think of their behavior as intelligent, think of them as having the capacity to behave intelligently, and even as having intelligence (though not conversational intelligence, of course).

My pet cube would behave in a way to encourage us to think of it in all these ways. When we come back to human beings, of course, we expect conversational intelligence, since we think of that as partly constitutive of distinctively human intelligence. But as Block defines 'intelligence' for the purpose of his argument, it is not obvious that it need involve conversational abilities, or indeed the possession of language. 'Intelligence' in his argument, he says, "means something like the possession of thought or reason" (1981 p.8). Later, he seems to suggest an even more modest reading of the term, as implying only the possession of genuine mental states or propositional attitudes (1981 p. 22, fn. 19). Unless we are to beg all sorts of controversial questions, we should not assume in this discussion that the possibility of testing an entity for conversational intelligence is a necessary condition for describing that entity as intelligent. We *can* do so on other grounds.

For human beings, as I have said, we regard a test of conversational intelligence as the best test of intelligence[12]. But in Block's argument it is conceded that the string-searcher could pass such a test, and the only question is whether, even so, we should still refuse to regard it as intelligent. What I have done is to try to weaken a certain intuition Block is appealing to in urging that we should, one that is based on a mix of considerations having to do with the *appearance* of autonomy, spontaneity and creativity. Whether the appearance of these in human beings is *mere* appearance or not, I have tried to show that systems very like Block's string-searcher machine could easily give such an appearance. I have suggested that if they did, we would characterize them as intelligent for same reasons we characterize ourselves as such, and that we would be right in doing so; that such a characterization would remain justified in *both* cases even if we found the appearance

[12]Though it may well break down unusual, Helen Keller-type, cases.

nothing more than that. This means that a system's having an internal mechanism that is incompatible with that appearance's being more than just appearance does not mean that it has an internal mechanism incompatible with having real intelligence. The heart of Block's appeal lies in a sort of *reductio*:

> If someone offered a definition of 'life' that had the unnoticed consequence that small stationary items such as paper clips are alive, one could refute him by pointing out the absurdity of the consequence, even if one had no very detailed account of what life really is with which to replace his. *In the same way*, one can refute the neo-Turing Test conception [of intelligence] by counter-example without having to say very much about what intelligence really is (1981 p. 28, my emphasis).

Block is right: one need not know much about life to know that a paper clip does not have any. It does not satisfy even the most obvious behavioral criteria we have for being a living thing. But Block's string-searcher, by his own stipulation, *does* satisfy our *behavioral* criteria for intelligence just as well as we do, if not better. I have tried to show that there is no compelling reason to think that it needs to satisfy any other condition, in particular, not one concerning its internal processing mechanisms, to be thought of as genuinely intelligent in the same sense as a human being[13].

BIBLIOGRAPHY

Biro, J. (1981). Persons as Corporate Entities and Corporations as Persons, *Nature and System* vol. 3.

Block, N. (1981). Psychologism and Behaviorism, *Philosophical Review* vol. 90 n. 1 (January), 5-43.

[13]I wish to thank Ray Elugardo, Kirk Ludwig and Chris Swoyer for help with various earlier attempts at formulating these arguments.

Dennett, D. (1971). Intentional Systems, *Journal of Philosophy* vol. 68 n. 4, reprinted in Dennett 1978b.

——. (1977). A Cure for the Common Code, *Mind* (April), reprinted in Dennett 1978b.

——. (1978a). Artificial Intelligence as Philosophy and as Psychology, in M. Ringle ed. *Philosophical Perspectives on Artificial Intelligence* (New York, Humanities Press) reprinted in Dennett 1978b.

——. (1978b). *Brainstorms: Philosophical Essays on Mind and Psychology* (Montgomery VT, Bradford Books).

Fodor, J. (1978). Tom Swift and his Procedural Grandmother, *Cognition* vol. 6 (May), 228-47.

—— (1980). Methodological Solipsism Considered as a Research Strategy in Cognitive Science, *The Behavioral and Brain Sciences* vol. 3, 63-109.

Searle, J. (1980) Minds, Brains and Programs, *The Behavioral and Brain Sciences* vol. 3, 417-424.

Weizenbaum, J. (1965). ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machine, *Communications of the Association for Computing Machinery* vol. 9, 36-45.

PUBLIUS:

THE JOURNAL OF FEDERALISM

Published by the
Center for the Study of Federalism
and University of North Texas

Editors: Daniel J. Elazar and John Kincaid

PUBLIUS is a quarterly journal now in its twenty-second year of publication. It is dedicated to the study of federal principles, institutions, and processes. PUBLIUS publishes articles, research notes, and book reviews on the theoretical and practical dimensions of the American federal system and intergovernmental relations and other federal systems throughout the world.

Forthcoming topical issues will feature articles on federalism and the Constitution, federalism and military rule in Nigeria, federalism and rights, counties in the federal system, federal preemption of state and local authority, federalism in Spain, and much more, as well as the PUBLIUS Annual Review of American Federalism edited by Ann O'M. Bowman and Michael A. Pagano.

ANNUAL SUBSCRIPTION RATES: Individual $25; Institutional $35. (Add $5.00 for foreign postage.)

Subscriptions and manuscript submissions should be sent to:

PUBLIUS: THE JOURNAL OF FEDERALISM
c/o Department of Political Science, University of North Texas
Denton, Texas 76203-5338 (817) 565-2313