The Philosophical Review, XC, No. 1 (January 1981)

PSYCHOLOGISM AND BEHAVIORISM

Ned Block

Let psychologism be the doctrine that whether behavior is intelligent behavior depends on the character of the internal information processing that produces it. More specifically, I mean psychologism to involve the doctrine that two systems could have actual and potential behavior *typical* of familiar intelligent beings, that the two systems could be exactly alike in their actual and potential behavior, and in their behavioral dispositions and capacities and counterfactual behavioral properties (i.e., what behaviors, behavioral dispositions, and behavioral capacities they would have exhibited had their stimuli differed)—the two systems could be alike in all these ways, yet there could be a difference in the information processing that mediates their stimuli and responses that determines that one is not at all intelligent while the other is fully intelligent.

This paper makes two claims: first, psychologism is true, and thus a natural behaviorist analysis of intelligence that is incompatible with psychologism is false. Second, the standard arguments against behaviorism are inadequate to defeat this natural behaviorist analysis of intelligence or to establish psychologism.

While psychologism is of course anathema to behaviorists,[1] it also seems wrong-headed to many philosophers who would *not* classify themselves as behaviorists. For example, Michael Dummett says:

> If a Martian could learn to speak a human language, or a robot be devised to behave in just the ways that are essential to a language speaker, an implicit knowledge of the correct theory of meaning for the language could be attributed to the Martian or the robot

[1] Indeed, Ryle's *The Concept of Mind* (London: Hutchinson, 1949) is a direct attack on psychologism. Ryle considers what we are judging "in judging that someone's performance is or is not intelligent," and he concludes: "Our inquiry is *not into causes* . . . but into capacities, skills, habits, liabilities and bents." See Jerry Fodor's *Psychological Explanation* (New York: Random House, 1968) for a penetrating critique of Ryle from a psychologistic point of view.

with as much right as to a human speaker, even though their internal mechanisms were entirely different.[2]

Dummett's view seems to be that what is relevant to the possession of a certain mental state is a matter of actual and potential behavior, and that internal processing is *not* relevant except to the extent that internal processing affects actual and potential behavior. I think that this Dummettian claim contains an important grain of truth, a grain that many philosophers wrongly take to be incompatible with psychologism.

This grain of truth can be elucidated as follows. Suppose we meet Martians, and find them to be behaviorally indistinguishable from humans. We learn their languages and they learn ours, and we develop deep commercial and cultural relations with them. We contribute to their journals and enjoy their movies, and vice versa. Then Martian and human psychologists compare notes, only to find that in underlying psychological mechanisms the Martians are very different from us. The Martian and human psychologists soon agree that the difference could be described as follows. Think of humans and Martians as if they were the products of conscious design. In any artificial intelligence project, there will be a range of engineering options. For example, suppose one wants to design a machine that makes inferences from information fed into it in the form of English sentences. One strategy would be to represent the information in the machine in English, and to formulate a set of inference rules that operate on English sentences. Another strategy would be to formulate a procedure for translating English into an artificial language whose sentences wear their logical forms on their faces. This strategy would simplify the inference rules, though at the computational cost of implementing the translation procedure. Suppose that the Martian and human psychologists agree that Martians and humans differ as if they were the products of a whole series of engineering decisions that differ along the lines illustrated. Should we conclude that the Martians are *not* intelligent after all? Obviously not! That would be crude human chauvinism. I suggest that philosophers reject psychologism in

[2] "What Is a Theory of Meaning (II)," in *Truth and Meaning*, ed. G. Evans and J. McDowell (London: Oxford University Press, 1976).

part because they (wrongly) see psychologism as involving this sort of chauvinism.

One of my purposes in this paper will be to show that psychologism does not in fact involve this sort of chauvinism.

If I succeed in showing psychologism to be true, I will have provided aid and comfort to those of us who have doubts about functionalism (the view that mental states are functional states—states definable in terms of their causal roles). Doubts about functionalism stem in part from the possibility of entities that look and act like people (and possess a network of internal states whose causal roles mirror those of our mental states), but differ from people in being operated by a network of homunculi whose aim is to simulate the functional organization of a person.[3] The presence of the homunculi can be used to argue that the homunculi-heads lack mentality. Defenders of functionalism are often inclined to "bite the bullet," replying along the following lines: "If I were to discover that my best friend and most valuable colleague was a homunculi-head, that should not lead me to regard him as lacking intelligence (or other aspects of mentality), since differences in internal goings-on that do not affect actual or potential behavior (or behavioral counterfactuals) are not relevant to intelligence." If this paper shows psychologism to be true, it blocks this line of defense of functionalism.

Let us begin the main line of argument by focusing on the well-known Turing Test. The Turing Test involves a machine in one room, and a person in another, each responding by teletype to remarks made by a human judge in a third room for some fixed period of time, e.g., an hour. The machine passes the test just in case the judge cannot tell which are the machine's answers and which are those of the person. Early perspectives on the

[3] See my "Troubles with Functionalism," in *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science, 9, ed. C. W. Savage (Minneapolis: University of Minnesota Press, 1978). Direct criticisms appear in William Lycan's "A New Lilliputian Argument against Machine Functionalism," *Philosophical Studies*, 35 (1979) and in Lycan's "Form, Function, and Feel," forthcoming in the *Journal of Philosophy*. See also Sydney Shoemaker's "Functionalism and Qualia," *Philosophical Studies*, 27 (1975), my reply, "Are Absent Qualia Impossible?" *Philosophical Review*, 89 (1980), and Shoemaker's rejoinder, "The Missing Absent Qualia Argument—a Reply to Block," forthcoming.

Turing Test reflected the contemporary view of what it was for something to be intelligent, namely that it act in a certain way, a way hard to define, but easy enough to recognize.

Note that the sense of "intelligent" deployed here—and generally in discussion of the Turing Test[4]—is *not* the sense in which we speak of one person being more intelligent than another. "Intelligence" in the sense deployed here means something like the possession of thought or reason.

One popular way of construing Turing's proposal is as a version of operationalism. "Being intelligent" is defined as passing the Turing Test, if it is administered (or alternatively, à la Carnap: if a system is given the Turing Test, then it is intelligent if and only if it passes). Construed operationally, the Turing Test conception of intelligence shares with other forms of operationalism the flaw of stipulating that a certain measuring instrument (the Turing Test) is *infallible*. According to the operationalist interpretation of the Turing Test as a definition of intelligence, it is absurd to ask of a device that passes the Turing Test whether it is *really* intelligent, and it is equally absurd to ask of a device that fails it whether it failed for some extraneous reason, but is nonetheless intelligent.

This difficulty can be avoided by going from the crude operationalist formulation to a familiar behavioral disposition formulation. On such a formulation, intelligence is identified not with the property of passing the test (if it is given), but rather with a behavioral *disposition* to pass the test (if it is given). On this behaviorist formulation, failing the Turing Test is not taken so seriously, since we can ask of a system that fails the test whether the failure *really does* indicate that the system lacks the disposition to pass the test. Further, passing the test is not *conclusive* evidence of a disposition to pass it, since, for example, the pass may have been accidental.

But the new formulation is nonetheless subject to deep difficulties. One obvious difficulty is its reliance on the discrimina-

[4] Turing himself said the question of whether the machine could *think* should "be replaced by" the question of whether it could pass the Turing Test, but much of the discussion of the Turing Test has been concerned with *intelligence* rather than thought. (Turing's paper [in *Mind*, 1950] was called "Computing Machinery and *Intelligence*" [emphasis added].)

tions of a human judge. Human judges may be able to discriminate *too well*—that is, they may be able to discriminate some *genuinely* intelligent machines from humans. Perhaps the responses of some intelligent machines will have a machinish style that a good human judge will be able to detect.

This problem could be avoided by altering the Turing Test so that the judge is not asked to say which is the machine, but rather is asked to say whether one or both of the respondents are, say, as intelligent as the average human. However, this modification introduces circularity, since "intelligence" is defined in terms of the judge's judgments of intelligence. Further, even ignoring the circularity problem, the modification is futile, since the difficulty just crops up in a different form: perhaps human judges will tend chauvinistically to regard some genuinely intelligent machines as unintelligent because of their machinish style of thought.

More importantly, human judges may be too easily fooled by mindless machines. This point is strikingly illustrated by a very simple program[5] (two hundred lines in BASIC), devised by Joseph Weizenbaum, which can imitate a psychiatrist by employing a small set of simple strategies. Its major technique is to look for key words such as "I," "you," "alike," "father," and "everybody." The words are ranked—for example, "father" is ranked above "everybody," and so if you type in "My father is afraid of everybody," the machine will respond with one of its "father" responses, such as "What else comes to mind when you think of your father?" If you type in "I know everybody laughed at me," you will get one of its responses to "everybody," for example, "Who in particular are you thinking of?" It also has techniques that simultaneously transform "you" into "I" and "me" into "you," so that if you type in "You don't agree with me," it can reply: "Why do you think that I don't agree with you?" It also *stores* sentences containing certain key words such as "my." If your *current* input contains no key words, but if you had earlier said "My boyfriend made me come here," it will

[5] "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the Association for Computing Machinery*, 9 (1965). See also M. Boden, *Artificial Intelligence* (New York: Basic Books, 1977).

"ignore" your current remark, saying instead, "Does that have anything to do with the fact that your boyfriend made you come here?" If all other tricks fail, it has a list of last ditch responses such as, "Who is the psychiatrist here, you or me?" Though this system is *totally* without intelligence, it proves *remarkably* good at fooling people in short conversations. Of course, Weizenbaum's machine rarely fools anyone for very long if the person has it in mind to explore the machine's capacities. But the program's extraordinary success (Weizenbaum's secretary asked him to leave the room in order to talk to the machine privately) reminds us that human gullibility being what it is, some more complex (but nonetheless unintelligent) program may be able to fool most any human judge. Further, since our tendency to be fooled by such programs seems dependent on our degree of suspicion, sophistication about machines, and other contingent factors, it seems silly to adopt a view of the nature of intelligence or thought that so closely ties it to human judgment. Could the issue of whether a machine *in fact* thinks or is intelligent depend on how gullible human interrogators tend to be?

In sum, human judges may be unfairly chauvinist in rejecting genuinely intelligent machines, and they may be overly liberal in accepting cleverly-engineered, mindless machines.

The problems just described could be avoided if we could specify in a non-question-begging way what it is for a sequence of responses to verbal stimuli to be a typical product of one or another style of intelligence. For then we would be able to avoid the dependence on human powers of discrimination that lies at the root of the problems of the last paragraph. Let us suppose, for the sake of argument, that we *can* do this, that is, that we can formulate a non-question-begging definition—indeed, a behavioristically acceptable definition—of what it is for a sequence of verbal outputs to be, as we shall say, "sensible," relative to a sequence of inputs. Though of course it is very doubtful that "sensible" can be defined in a non-question-begging way, it will pay us to *suppose* it can, for as we shall see, even such a definition would not save the Turing Test conception of intelligence.

The role of the judge in Turing's definition of intelligence is to avoid the problem of actually specifying the behavior or behavioral dispositions thought to constitute intelligence.

Hence my supposition that "sensible" can be defined in a non-question-begging way amounts to the suggestion that we ignore one of the usual criticisms of behaviorists—that they cannot specify their behavioral dispositions in a non-question-begging way. This is indeed an enormous concession to behaviorism, but it will not play an important role in what follows.

We can now propose a version of the Turing Test conception of intelligence that avoids the problems described:

> Intelligence (or more accurately, conversational intelligence) is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be.

The point of the "whatever they may be" is to emphasize that this account avoids relying on anyone's ability to come up with clever questions; for in order to be intelligent according to the above-described conception, the system must be disposed to respond sensibly not only to what the interlocutor *actually* says, but to whatever he *might* have said as well.

While the definition just given is a vast improvement (assuming that "sensible" can be adequately defined), it is still a clearly behaviorist formulation. Let us now review the standard arguments against behaviorism with an eye towards determining whether the Turing Test conception of intelligence is vanquished by them.

Probably the most influential argument against behaviorism is due to Chisholm and Geach.[6] Suppose a behaviorist analyzes someone's wanting an ice cream cone as his having a set of behavioral dispositions such as the disposition to grasp an ice cream cone if one is "presented" to him. But someone who wants an ice cream cone will be disposed to grasp it only if he *knows* it is an ice cream cone (and not in general if he thinks it is a tube of axle grease being offered to him as a joke) and only if he does not *believe* that taking an ice cream cone would conflict with other *desires* of more importance to him (for example, the desire to

[6] See Roderick Chisholm, *Perceiving* (Ithaca, N. Y.: Cornell University Press, 1957), ch. 11, and Peter-Geach, *Mental Acts* (London: Routledge, 1957), p. 8.

avoid an obligation to return the favor). In short, which behavioral dispositions a desire issues in depends on the *other* mental states of the desirer. And similar points apply to behaviorist analyses of belief and of many other mental states. Conclusion: one cannot define the conditions under which a given mental state will issue in a given behavioral disposition without adverting to *other mental states.*

Another standard argument against behaviorism flows out of the Chisholm-Geach point. If a person's behavioral dispositions depend on a *group* of mental states, perhaps *different* mental groups can produce the *same* behavioral dispositions. This line of thought gave rise to the "perfect actor" family of counterexamples. As Putnam[7] argued in convincing detail, it is possible to imagine a community of perfect actors (Putnam's super-super-spartans) who, in virtue of lawlike regularities, lack the behavioral dispositions envisioned by the behaviorists to be associated with pain, even though they do in fact have pain. This shows that no behavioral disposition is necessary for pain, and an exactly analogous example of perfect pain-pretenders shows that no behavioral disposition is sufficient for pain either.

Another less important type of traditional counterexample to behaviorism is illustrated by paralytics and brains in vats. Like Putnam's super-super-spartans, they can have pain without the usual dispositions.

When I speak of the "standard objections to behaviorism" in what follows, I shall have these three types of objection in mind: the Chisholm-Geach objection, the perfect actor objection, and the objection based on paralytics and the like.[8]

---

[7] "Brains and Behavior," in Putnam's collected papers (Volume II), *Mind, Language and Reality* (London: Cambridge University Press, 1975).

[8] While the Chisholm-Geach objection and the perfect actor objection ought in my view to be considered the main objections to behaviorism in the literature, they are not on *everybody's* list. Rorty, for example (*Philosophy and the Mirror of Nature* [Princeton, N. J.: Princeton University Press, 1979]), has his own list (p. 98). Rorty and others make heavy weather of one common objection that I have ignored: that behaviorism's analyses of mental states are supposed to be analytic or true in virtue of the meanings of the mental terms. I have ignored analyticity objections in part because behaviorism's main competitors, physicalism and functionalism, are often held in versions that involve commitment to analytic truth (for example, by Lewis and Shoemaker). Further, many behaviorists have been willing to settle for conceptual connections

I. Do the Standard Objections to Behaviorism Dispose of Behaviorist Conceptions of Intelligence?

The three arguments just reviewed are generally and rightly regarded as decisive refutations of behaviorist analyses of many mental states, such as belief, desire, and pain. Further, they serve to refute one quite plausible behaviorist analysis of intelligence. Intelligence is plausibly regarded as a second order mental property, a property that consists in having first order mental states—beliefs, desires, etc.—that are caused to change in certain ways by changes in one another and in sensory inputs. If intelligence is indeed such a second order property, and given that the behaviorist analyses of the first order states are false, one can conclude that a plausible behaviorist view of intelligence is false as well.[9]

But it would be unfair to behaviorism to leave the matter here. Behaviorists generally mean their dispositions to be "pure dispositions." Ryle, for example, emphasized that "to possess a dispositional property is not to be in a particular state or to undergo a particular change."[10] Brittleness, according to Ryle, is not a *cause* of breaking, but merely *the fact* of breaking easily. Similarly, to attribute pain to a person is not to attribute a cause or effect of anything, but simply to say *what he would do* in certain circumstances. However, the notion just mentioned of intelligence as a second order property is at its most plausible when first order mental states are thought of as entities that *have causal roles*. Since pure dispositions do not have causal roles in any straightforward sense, the analysis of intelligence as a second order property should seem unsatisfactory to a behaviorist, even if it is the right analysis of intelligence. Perhaps this explains why behaviorists and behaviorist-sympathizers do not seem to have adopted a view of intelligence as a second order property.

Secondly, an analysis of intelligence along roughly the lines indicated in what I called the Turing Test conception of intel-

"weaker" than analyticity, and I see no point in exploring such weakened versions of the thesis when behaviorism can be refuted quite independently of the analyticity issue.

[9] I am indebted here to Sydney Shoemaker.

[10] Op. cit., p. 43.

ligence is natural for the behaviorist because it arises by patching a widely known operationalist formulation. It is not surprising that such a position is popular in artificial intelligence circles.[11] Further, it seems to be regarded sympathetically by many philosophers who accept the standard arguments against behaviorist analyses of beliefs, desires, etc.[12]

Another attraction of an analysis along the lines suggested by the Turing Test conception of intelligence is that such an analysis can *escape the standard objections to behaviorism.* If I am right about this, then it would certainly be foolish for the critic of behaviorism to regard behaviorism with respect to intelligence as obliterated by the standard objections, ignoring analyses along the lines of the Turing Test conception of intelligence. For these reasons, I will now return to an examination of how well the Turing Test conception of intelligence fares when faced with the standard objections.

The Turing Test conception of intelligence offers a necessary and sufficient condition of intelligence. The standard objections are effective against the necessary condition claim, but not against the sufficient condition claim. Consider, for example, Putnam's perfect actor argument. The super-super-spartans have pain, though they have no disposition to pain behavior. Similarly, a machine might *be* intelligent, but not be disposed to *act* intelligently because, for example, it might be programmed to believe

---

[11] See R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures* (Hillsdale, N. J.: Lawrence Erlbaum Assoc., 1977). See also Weizenbaum's description of the reaction to his ELIZA program in his *Computer Power and Human Reason* (San Francisco: Freeman, 1976).

[12] There is, admittedly, something odd about accepting a behaviorist analysis of intelligence while rejecting (on the standard grounds) behaviorist theories of belief, desire, etc. Dennett's view, as I understand it, comes close to this (see note 29), though the matter is complicated by Dennett's skepticism about many first order mental states. (See *Brainstorms* [Montgomery: Bradford, 1978], especially the Introduction, and Dennett's support of Ryle against Fodor's psychologism—p. 96 of *Brainstorms*. See also Dennett's Mary-Ruth-Sally parable on p. 105 of *Brainstorms*.) In discussions among computer scientists who accept something like the Turing Test conception, the "oddness" of the position doesn't come to the fore because these practitioners are simply not *interested* in making machines that believe, desire, feel, etc. Rather, they focus on machines that are intelligent in being able to reason, solve problems, etc.

that acting intelligently is not *in its interest*. But what about the converse sort of case? A perfect actor who pretends to *have* pain seems as plausible as the super-super-spartans who pretend to *lack* pain, but *this* sort of perfect actor case does *not* seem to transfer to intelligence. For how could an unintelligent system perfectly pretend to be intelligent? It would seem that any system that is *that* good at pretending to be intelligent would have to *be* intelligent. So no behavioral disposition is necessary for intelligence, but *as far as this standard objection is concerned,* a behavioral disposition may yet be sufficient for intelligence. A similar point applies with respect to the Chisholm-Geach objection. The Chisholm-Geach objection tells us that a disposition to pain behavior is not a sufficient condition of having pain, since the behavioral disposition could be produced by a number of different combinations of mental states, e.g., [pain + a normal preference function] or by [no pain + an overwhelming desire to appear to have pain]. Turning to intelligent behavior, we see that it normally is produced by [intelligence + a normal preference function]. But could intelligent behavior be produced by [no intelligence + an overwhelming desire to appear intelligent]? Indeed, could there be *any* combination of mental states and properties *not including intelligence* that produces a lawful and thoroughgoing disposition to act intelligently? It seems not. So it seems that the Chisholm-Geach objection does not refute the claim of the Turing Test conception that a certain disposition is sufficient for intelligence.

Finally, the standard paralytic and brain in the vat examples are only intended to apply to claims of necessary conditions—not sufficient conditions—of mental states.

The defect I have just pointed out in the case against the behaviorist view of intelligence is a moderately serious one, since behaviorists have tended to focus on giving sufficient conditions for the application of mental terms (perhaps in part because of their emphasis on the connection between the meaning of "pain" and the circumstances in which we learned to apply it). Turing, for example, was willing to settle for a "sufficient condition" formulation of his behaviorist definition

of intelligence.[13] One of my purposes in this paper is to remedy this defect in the standard objections to behaviorism by showing that no behavioral disposition is sufficient for intelligence.

I have just argued that the standard objections to behaviorism are only partly effective against the Turing Test conception of intelligence. I shall now go one step further, arguing that there is a reformulation of the Turing Test conception of intelligence that avoids the standard objections *altogether*. The reformulation is this: substitute the term "capacity" for the term "disposition" in the earlier formulation. As mentioned earlier, there are all sorts of reasons why an intelligent system may fail to be disposed to act intelligently: believing that acting intelligently is not in its interest, paralysis, etc. But intelligent systems that do not want to act intelligently or are paralyzed still have *the capacity* to act intelligently, even if they do not or cannot exercise this capacity.

Let me say a bit more about the difference between a behavioral disposition and a behavioral capacity. A capacity to ϕ need not result in a disposition to ϕ unless certain *internal* conditions are satisfied—say, the appropriate views or motivation or not having curare in one's bloodstream. To a first approximation, a disposition can be specified by a set (perhaps infinite) of input-output conditionals.

If i_1 obtains, then o_1 is emitted
If i_2 obtains, then o_2 is emitted
and so on.[14]

---

[13] Turing says:

> The game may perhaps be criticized on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection. [op. cit., p. 435]

[14] A disposition to ϕ would be more revealingly described in terms of conditionals all of whose consequents are "ϕ is emitted." But in the cases of the "pain behavior" or "intelligent behavior" of interest to the behaviorist, what output is appropriate depends on the input.

A corresponding first stab at a specification of a capacity, on the other hand, would involve mentioning *internal states* in the antecedents of the conditionals.

> If s_a and i_a obtain, then o_a is emitted
> If s_b and i_b obtain, then o_b is emitted
> and so on,

where s_a and s_b are internal states.[15] In humans, such states would include beliefs and desires and working input and output organs at a minimum, though a machine could have a capacity the exercise of which is contingent only on nonmental internal parameters, e.g., whether its fuses are intact.

What I have said about the difference between a disposition and a capacity is very sketchy, and clarification is needed, especially with regard to the question of what sorts of internal states are to be specified in the antecedents of the conditionals. If paralytics are to be regarded as possessing behavioral capacities, these internal states will have to include specifications of functioning input and output devices. And if the systems that believe that acting intelligently is not in their interest are to have the required capacity, internal states will have to be specified such that if they *were* to obtain, the system would believe that acting intelligently *is* in its interest. Notice, however, that the behaviorist need not be committed to these *mentalistic descriptions*

---

[15] Of the inadequacies of this sort of analysis of dispositions and capacities of which I am aware, the chief one is that it seems implausible that in attributing a disposition or a capacity, one commits oneself to an infinite (or even a very large) number of specific conditionals. Rather, it seems that in saying that x has the capacity to ϕ, one is saying something *quite vague* about the sort of internal and external conditions in which x would ϕ. Notice, however, that it won't do to be *completely* vague, to analyze "x has the capacity to ϕ" as "possibly, x ϕs," using a notion of possibility that holds entirely unspecified features of the actual world constant. For such an analysis would commit its proponents to ascribing too many capacities. For example, since there is a possible world in which Jimmy Carter has had a womb and associated paraphernalia surgically inserted, Jimmy Carter (*the actual one*) would have the capacity to bear children. There is a difference between the capacities someone *has* and the capacities he *might have had*, and the analysis of "x has the capacity to ϕ" as "possibly, x ϕs" does not respect this distinction.

of the internal states; physiological or functional descriptions will do.[16]

The reader may suspect that the reformulation of behaviorism in terms of capacities that I have suggested avoids the standard objections to behaviorism only because it concedes *too much*. The references to internal states—even under physiological or functional descriptions—may be seen as too great a concession to psychologism (or other nonbehavioristic doctrines) for any genuine behaviorist to make. I reply: so much the better for my purposes, for I intend to show that this concession is not *enough*, and that the move from behavioral dispositions to behavioral capacities *will not save behaviorism*.

I now propose the reformulation suggested by the preceding remarks; let us call it the *neo-Turing Test conception of intelligence*.

> Intelligence (or, more accurately, conversational intelligence) is the capacity to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be.

Let us briefly consider the standard objections to behaviorism in order to show that the neo-Turing Test conception avoids them. First, intelligent paralytics and brains in vats provide no counterexample, since they do have the capacity to respond sensibly, though they lack the means to exercise the capacity. Second, consider the "perfect actor" objection. An intelligent being who perfectly feigns stupidity nonetheless has the capacity to respond sensibly. Further, as in the disposition case, it would seem that no one could have the capacity to pretend perfectly to be intelligent without actually being intelligent. Third, the new formulation entirely disarms the Chisholm-Geach objection. There are many combinations of beliefs and desires that could cause an intelligent being to fail to be *disposed* to respond sensibly, but these beliefs and desires would not destroy the being's *capac-*

[16] The departure from behaviorism involved in appealing to internal states, physiologically or functionally described, is mitigated somewhat when the point of the previous footnote is taken into account. The physiological/functional descriptions in a *proper* analysis of capacities may be so vague as to retain the behavioristic flavor of the doctrine.

ity to respond sensibly. Further, as I have mentioned repeatedly, it is hard to see how any combination of mental states not including intelligence could result in the capacity to respond in an intelligent manner to arbitrary sequences of stimuli.

One final point. Notice that my concession that "sensible" can be defined in a behavioristically adequate way is *not* what is responsible for the fact that the neo-Turing Test conception of intelligence evades the standard objections. What does the job is first the difficulty of conceiving of someone who can pretend perfectly to be intelligent without actually being intelligent, and second, the move from dispositions to capacities.

II. The Argument for Psychologism and Against Behaviorism

My strategy will be to describe a machine that produces (and thus has the capacity to produce) a sensible sequence of verbal responses to verbal stimuli. The machine is thus intelligent according to the neo-Turing Test conception of intelligence (and also according to the cruder versions of this conception). However, according to me, a knowledge of the machine's internal information processing shows conclusively that it is totally lacking in intelligence.

I shall now describe my unintelligent machine. 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 (indeed, a finite number of typable letter strings), 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 in the sense described above. 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. We need not be very restrictive as to what is to count as sensible. For example, if sentence 1 is "Let's see you talk nonsense," it would be sensible for sentence 2 to be nonsensical. The set of sensible strings so defined is a finite set that could in principle be listed by a very large and clever team working for a long time, with a very large grant and a lot of mechanical help, *exercising imagination and judgment* about what is to count as a sensible string.

Presumably the programmers will find that in order to produce really convincing sensible strings, they will have to think of themselves as simulating some definite personality with some definite history. They might choose to give the responses my Aunt Bertha might give if she were brought to a room with a teletype by her errant nephew and asked to answer "silly" questions for a time.

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.[17]

The reader may be helped by seeing a variant of this machine in which the notion of a sensible string is replaced by the notion of a sensible branch of a tree structure. Suppose the interrogator goes first, typing in one of $A_1 \ldots A_n$. The programmers produce *one* sensible response to each of these sentences, $B_1 \ldots B_n$. For each of $B_1 \ldots B_n$ the interrogator can make various replies, so many branches will sprout below each of $B_1 \ldots B_n$. Again, for each of these replies, the programmers produce one sensible response, and so on. In this version of the machine, all the X-initial strings can be replaced by a single tree with a single token of X as the head node; all the XYZ-initial strings can be replaced by a branch of that tree with Y and Z as the next nodes, and so forth. This machine is a tree-searcher instead of a string-searcher.

[17] A version of this machine was sketched in my "Troubles with Functionalism," op. cit.

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.* Note also that its limitation to Turing Tests of an hour's length is not essential. For a Turing Test of *any* given length, the machine could in principle be programmed in just the same way to pass a Turing Test of that length.

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 (along with the older and cruder Turing Test conceptions). 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.

I haven't shown *quite* what I advertised initially, since I haven't shown that the machine could duplicate the response properties of a real person. But what I have shown is close enough for me, and besides, it doesn't change the essential point of the example if we imagine the programmers deciding *exactly* what Aunt Bertha would say on the basis of a psychological or physiological theory of Aunt Bertha.

We can now see why psychologism is not incompatible with the point made earlier in connection with the Martian example. The Martian example suggested that it was doubtful that there would be any single natural kind of information processing that must be involved in the production of all intelligent behavior. (I argued that it would be chauvinist to refuse to classify Martians as intelligent *merely* because their internal information processing is very different from ours.) Psychologism is not chauvinist because psychologism requires only that intelligent behavior *not* be the product of a (at least one) certain kind of internal processing. One can insist that behavior which has a certain etiology cannot be intelligent behavior without holding that all intelligent behavior must have the same "kind" of etiology.

The point of the machine example may be illuminated by comparing it with a two-way radio. If one is speaking to an intelligent person over a two-way radio, the radio will normally emit sensible replies to whatever one says. But the radio does not do this in virtue of a capacity to make sensible replies that *it* possesses. The two-way radio is like my machine in being a *conduit* for intelligence, but the two devices differ in that my machine has a crucial capacity that the two-way radio lacks. In my machine, no causal signals from the interrogators reach those who think up the responses, but in the case of the two-way radio, the person who thinks up the responses has to hear the questions. In the case of my machine, the causal efficacy of the programmers is limited to what they have stored in the machine before the interrogator begins.

The reader should also note that my example is really an extension of the traditional perfect pretender counterexample, since the machine "pretends" to be intelligent without actually being intelligent. Once one notes this, it is easy to see that a *person* could have a capacity to respond intelligently, even though the intelligence he exhibits is not *his*—for example, if he memorizes responses in Chinese though he understands only English.[18] An idiot with a photographic memory, such as Luria's famous mnemonist, could carry on a brilliant philosophical conversation if provided with strings by a team of brilliant philosophers.[19]

[18] This sort of point is discussed in somewhat more detail at the end of the paper.

[19] What I say here should not be taken as indicating that the standard objections *really do* vanquish the neo-Turing Test conception of intelligence after all. If the idiot can be said to have the mental state [no intelligence + an overwhelming desire to appear intelligent], the sense of "intelligence" used is the "comparative" sense, not the sense we have been concerned with here (the sense in which intelligence is the possession of thought or reason) If the idiot *wants* to appear intelligent (in the comparative sense) and *thinks* that he can do so by memorizing strings, then he *is* intelligent in the sense of possessing (at least minimally) thought or reason.

Whether one thinks my objection is really just a variant of the "perfect actor" objection depends on how closely one associates the perfect actor objection with the Chisholm-Geach objection. If we associate the perfect actor objection quite closely with the Chisholm-Geach objection, as I think is historically accurate (see p. 324 of Putnam's *Mind, Language and Reality),* then we will take the point of the perfect actor objection to be that different *groups of mental states* can produce the same behavioral dispositions. [mental state

The machine, as I have described it thus far, is limited to typewritten inputs and outputs. *But this limitation is inessential, and that is what makes my argument relevant to a full-blooded behaviorist theory of intelligence*, not just to a theory of conversational intelligence. What I need to show to make my point is that the kind of finiteness assumption that holds with respect to typewritten inputs and outputs also holds with respect to the whole range of sensory stimulation and behavior. If I can show this, then I can generalize the idea of the machine I described to an unintelligent robot that nonetheless acts in every possible situation just as intelligently as a person.

The sort of finiteness claim that I need can be justified both empirically and conceptually. The empirical justifications are far too complex to present here, so I will only mention them briefly. First, I would claim that enough is now known about sensory physiology to back up the assertion that every stimulus parameter that is not already "quantized" could be quantized without making any difference with respect to effects on the brain or on behavior, provided that the "grain" of quantization is fine enough. Suppose that all of your sense organs were covered by a surface that effected an "analog-to-digital conversion." For example, if some stimulus parameter had a value of .111 . . . units, the surface might change it to .11 units. Provided that the grain was fine enough (not too many decimal places are "lopped off"), the analog-to-digital conversion would make no mental or behavioral difference. If this is right, then one could take the output of the analog-digital converter as the relevant stimulus, and so there would be a finite number of possible sequences of arrays of stimuli in a finite time.

I am told that a similar conclusion can actually be reached with respect to *any* physical system that can be regarded as having inputs and outputs. The crucial claim here is that no physical system could be an infinitely powerful amplifier, so given a "power of amplification," one could impose a corresponding quantiza-

---

x + a normal preference function] can produce the same behavioral disposition as [lack of mental state x + a preference function that gives infinite weight to seeming to have mental state x]. My machine is *not a perfect actor in this sense,* since it has no mental states, and hence no groups of mental states either.

tion of the inputs that would not affect the outputs. I don't know enough physics to pursue this line further, so I won't.

The line of argument for my conclusion that I want to rely on is more conceptual than empirical. The point is that our *concept* of intelligence allows an intelligent being to have quantized sensory devices. Suppose, for example, that Martian eyes are like movie cameras in that the information that they pass on to the Martian brain amounts to a series of newspaper-like "dot" pictures, i.e., matrices containing a large number of cells, each of which can be either black or white. (Martians are color-blind.) If Martians are strikingly like us in appearance, action, and even internal information processing, no one ought to regard their movie camera eyes (and other finitary sense organs) as showing they are not intelligent. However, note that since there are a finite number of such "dot" pictures of a given grain, there are a finite number of *sequences* of such pictures of a given duration, and thus a finite number of *possible visual stimuli* of a given duration.

It is easy to see that both the empirical and the conceptual points support the claim that an intelligent being could have a finite number of possible sequences of types of stimuli in a finite time (and the same is also true of responses). But then the stimulus sequences could in principle be catalogued by programmers, just as can the interrogator's remarks in the machine described earlier. Thus, a robot programmed along the lines of the machine I described earlier could be given every behavioral capacity possessed by humans, via a method of the sort I have already described. In sum, while my remarks so far have dealt mainly with a behaviorist account of *conversational* intelligence, broadening the argument to cover a behaviorist theory of intelligence *simpliciter* would raise no new issues of principle. In what follows, I shall return for convenience to a discussion of conversational intelligence.

By this time, the reader may have a number of objections. Given the heavy use of the phrase "in principle" above, you may feel that what this latest wrinkle shows is that the sense of "in principle possible" in which *any* of the machines I described are

in principle possible is a bit strange. Or you may object: "Your machine's capacity to pass the Turing Test *does* depend on an arbitrary time limit to the test." Or: "You are just stipulating a new meaning for the word 'intelligent.' " Or you may want to know what I would say if *I* turned out to be one.

I will now attempt to answer these and other objections. If an objection has a subscripted numeral (e.g., 3a), then it depends on the immediately preceding objection or reply. However, the reader can skip any other objection or reply without loss of continuity.

Objection 1. Your argument is too strong in that it could be said of *any* intelligent machine that the intelligence it exhibits is that of its programmers.

Reply. I do *not* claim that the intelligence of *every* machine designed by intelligent beings is merely the intelligence of the designers, and no such principle is used in my argument. If we ever do make an intelligent machine, presumably we will do it by equipping it with mechanisms for learning, solving problems, etc. Perhaps we will find general principles of learning, general principles of problem solving, etc., which we can build into it. But though we *make* the machine intelligent, the intelligence it exhibits is *its* own, just as our intelligence is no less ours, even if it was produced mainly by the enormously skillful efforts of our parents.

By contrast, if my string-searching machine emits a clever pun P, in response to a conversation C, then the sequence CP is literally one that was thought of and included by the programmers. Perhaps the programmers will say of one of their colleagues, "Oh, Jones thought of that pun—he is so clever."

The trouble with the neo-Turing Test conception of intelligence (and its predecessors) 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.* As I suggested, only a partly etiological notion of intelligent behavior will do the trick.

Objection 2. If the strings were recorded before this year, the machine would not respond the way a person would to a sentence like "What do you think of the latest events in the Mid-East?"

Reply. A system can be intelligent, yet have no knowledge of current events. Likewise, a machine can *imitate* intelligence without *imitating* knowledge of current events. The programmers could, if they liked, choose to simulate an intelligent Robinson Crusoe who knows nothing of the last twenty-five years. Alternatively, they could undertake the much more difficult task of reprogramming periodically to simulate knowledge of current events.

Objection 3. You have argued that a machine with a certain internal mechanical structure is not intelligent, even though it seems intelligent in every *external* respect (that is, in every external respect examined in the Turing Test). But by introducing this internal condition, aren't you in effect merely suggesting a linguistic stipulation, a new meaning for the word "intelligent"? We *normally* regard input-output capacities as criterial for intelligence. All you are doing is suggesting that we adopt a new practice, involving a *new* criterion which includes something about what goes on inside.

Reply. Jones plays brilliant chess against two of the world's foremost grandmasters at once. You think him a genius until you find out that his method is as follows. He goes second against grandmaster G_1 and first against G_2. He notes G_1's first move against him, and then makes the same move against G_2. He awaits G_2's response, and makes the same move against G_1, and so on. Since Jones's method itself was one he read about in a comic book, Jones's performance is no evidence of his intelligence. As this example[20] illustrates, it is a feature of our concept of intelligence, that to the degree that a system's performance merely echoes the intelligence of another system, the first system's performance is thereby misleading as an indication of its in-

[20] Such examples were suggested by Dick Boyd and Georges Rey in their comments on an earlier rendition of this paper. Rey tells me the chess story is a true tale.

telligence. Since my machine's performances are *all* echoes, these performances provide no reason to attribute intelligence to it.[21]

The point is that though we *normally* ascertain the intelligence of a system by trying to assess its input-output capacities, it is part of our ordinary concept of intelligence that input-output capacities can be misleading. As Putnam has suggested, it is part of the logic of natural kind terms that what seems to be a stereotypical X can turn out not to be an X at all if it fails to belong to the same scientific natural kind as the main body of things we have referred to as X's.[22] If Putnam is right about this, one can never accuse someone of "changing the meaning" of a natural kind term *merely* on the ground that he says that something that satisfies the standard "criteria" for X's is not an X.

Objection 3a. I am very suspicious of your reply to the last objection, especially your introduction of the Putnam point. Is it not rather chauvinist to suppose that a system has to be scientifically *like us* to be intelligent? Maybe a system with information processing very unlike ours does not belong in the extension of our term "intelligence"; but it is equally true that we do not belong in the extension of *its* term "shmintelligence." And who is to say that intelligence is any better than shmintelligence?

Reply. I have not argued that the *mere fact* of an information processing *difference* between my machine and us cuts any ice.

[21] The reader should not conclude from the "echo" examples that what *makes* my machine unintelligent is that its responses are echoes. Actually, what makes it unintelligent is that its responses are *mere* echoes, i.e., its information processing is of the most elementary sort (and the appearances to the contrary are merely the echoes of genuinely intelligent beings). Notice that such a machine would be just as unintelligent if it were produced by a cosmic accident rather than by the long creative labors of intelligent people. What makes this accidentally produced machine unintelligent is, as before, that its information processing is of the most elementary sort; the appearances to the contrary are produced in this case not via echoes, but by a cosmic accident.

[22] Hilary Putnam, "The Meaning of 'Meaning,'" in *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, 7, ed. Keith Gunderson (Minneapolis: University of Minnesota Press, 1975). See also Saul Kripke, "Naming and Necessity," in *Semantics and Natural Language*, ed. G. Harman and D. Davidson (Dordrecht, Holland: Reidel, 1972).

Rather, my point is based on the *sort* of information processing difference that exists. My machine lacks the kind of "richness" of information processing requisite for intelligence. Perhaps this richness has something to do with the application of abstract principles of problem solving, learning, etc. I wish I could say more about just what this sort of richness comes to. But I have chosen a much less ambitious task: to give a clear case of something that *lacks* that richness, but nonetheless behaves as if it were intelligent. If someone offered a definition of "life" that had the unnoticed consequence that small stationery 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 by counterexample without having to say very much about what intelligence really is.

Objection 4. Suppose it turns out that human beings, including you, process information in just the way that your machine does. Would you insist that humans are not intelligent?

Reply. I'm not very sure of what I would say about human intelligence were someone to convince me that human information processing is the same as that of my machine. However, I do not see that there is any *clearly and obviously correct* response to this question against which the responses natural for someone with my position can be measured. Further, none of the more plausible responses that I can think of are incompatible with what I have said so far.

Assume, for example, a theory of reference that dictates that in virtue of the causal relation between the word "intelligence" and human information processing, human information processing is intelligent *whatever* its nature.[23] Then, if I were con-

[23] The theory sketched in Putnam, op. cit., might be taken to have this consequence. Whether it does have this consequence depends on whether it dictates that there is *no* descriptive component at all to the determination of the reference of natural kind terms. It seems certain that there is *some* descriptive component to the determination of the reference of natural kind terms, just as there is some descriptive component to the determination of the reference of names. There is a possible world in which Moses was an Egyptian

vinced that humans process information in the manner of my machine, I should admit that my machine is intelligent. But how is this incompatible with my claim that my machine is not *in fact* intelligent? Tweaking me with "What if *you* turned out to be one?" is a bit like tweaking an atheist with "What if you turned out to be God?" The atheist would have to admit that if he were God, then God would exist. But the atheist could concede this counterfactual without giving up atheism. If the word "intelligence" is firmly anchored to human information processing, as suggested above, then my position is committed to the *empirical claim* that human information processing is not like that of my machine. But this is a perfectly congenial claim, one that is supported both by common sense and by empirical research in cognitive psychology.

Objection 5. You keep insisting that we do not process information in the manner of your machine. What makes you so sure?

Reply. I don't see how someone could make such an objection without being somewhat facetious. You will have no difficulty coming up with responses to my arguments. Are we to take seriously the idea that someone long ago recorded both what I said and a response to it and inserted both in your brain? Common sense recoils from such patent nonsense. Further, pick any issue of any cognitive psychology journal, and you will see attempts at experimental investigation of our information processing mechanisms. Despite the crudity of the evidence, it tells overwhelmingly against the string-searching idea.

---

fig merchant who spread tall tales about himself, but is there a possible world in which Moses was a brick? Similarly, even if there is a possible world in which tigers are automata, is there a possible world in which tigers exist, but are ideas? I would argue, along these lines, that the word "intelligence" attaches to whatever natural kind our information processing belongs to (assuming it belongs to a single natural kind) *unless* our information processing fails the minimal descriptive requirement for intelligence (as ideas fail the minimal descriptive requirement for being tigers). String-searchers, I would argue, *do* fail to have the minimal requirement for intelligence.

Our cognitive processes are undoubtedly much more mechanical than some people like to think. But there is a vast gap between our being more mechanical than some people like to think and our being a machine of the sort I described.

Objection 6. Combinatorial explosion makes your machine impossible. George Miller long ago estimated[24] that there are on the order of 10^{30} grammatical sentences 20 words in length. Suppose (utterly arbitrarily) that of these 10^{15} are semantically well formed as well. An hour-long Turing Test would require perhaps 100 such sentences. That makes 10^{1500} strings, a number which is greater than the number of particles in the universe.

Reply. My argument requires only that the machine be *logically* possible, not that it be feasible or even nomologically possible. Behaviorist analyses were generally presented as *conceptual analyses*, and it is difficult to see how conceptions such as the neo-Turing Test conception could be seen in a very different light. Could it be an *empirical hypothesis* that intelligence is the capacity to emit sensible sequences of outputs relative to input sequences? What sort of empirical evidence (other than evidence from *linguistics*) could there be in favor of such a claim? If the neo-Turing Test conception of intelligence is seen as something on the order of a claim about the concept of intelligence, then the mere *logical* possibility of an unintelligent system that has the capacity to pass the Turing Test is enough to refute the neo-Turing Test conception.

It may be replied that although the neo-Turing Test conception clearly is not a *straightforwardly* empirical hypothesis, still it may be *quasi-empirical*. For it may be held that the identification of intelligence with the capacity to emit sensible output sequences is a *background* principle or law of empirical psychology. Or it may be offered as a rational reconstruction (of our vague common sense conception of intelligence) which will be fruitful

[24] G. Miller, E. Galanter, and K. Pribram, *Plans and the Structure of Behavior* (New York: Holt, 1960), p. 146.

in future empirical psychological theories. In both cases, while no empirical evidence could *directly* support the neo-Turing Test conception, still it could be held to be part of a perspective that could be empirically supported as a whole.[25]

This reply would carry some weight if any proponent of the neo-Turing Test conception had offered the *slightest reason* for thinking that such a conception of intelligence is likely to contribute to the fruitfulness of empirical theories that contain it. In the absence of such a reason (and, moreover, in the presence of examples that suggest the contrary—behaviorist psychology and Turingish approaches to artificial intelligence—see footnote 11), why should we take the neo-Turing Test conception seriously as a quasi-empirical claim?

While this reply suffices, I shall add that my machine may indeed be nomologically possible. Nothing in contemporary physics prohibits the possibility of matter in some part of the

---

[25] What follows is one rejoinder for which I only have space for a brief sketch. If intelligence = sensible response capacity (and if the terms flanking the "=" are rigid), then the *metaphysical* possibility of my machine is enough to defeat the neo-Turing Test conception, even if it is not nomologically possible. (The claim that there are metaphysical possibilities that are not also nomological possibilities is one that I cannot argue for here.)

What if the neo-Turing Test conception of intelligence is formulated not as an identity claim, but as the claim that a certain capacity is nomologically necessary and sufficient for intelligence? I would argue that if F is nomologically necessary and sufficient for G, then one of the following holds:

(a) This nomological coextensivity is an ultimate law of nature.

(b) This nomological coextensivity can be explained in terms of an underlying mechanism.

(c) $F = G$.

In case (c), the claim is vulnerable to the point of the previous paragraph. Case (a) is obviously wrong. And in case (b), intelligence must be identifiable with something other than the capacity to give sensible responses. Suppose, for example, that we can give a mechanistic account of the correlation of intelligence with sensible response capacity by showing that intelligence requires a certain sort of cognitive structure, and creatures with such a cognitive structure have the required capacity. But then intelligence should be identified with *the cognitive structure* and not with the capacity. See my "Reductionism," in the *Encyclopedia of Bioethics* (New York: Macmillan, 1978), for a brief discussion of some of these ideas.

universe that is infinitely divisible. Indeed, whenever the latest "elementary" particle turns out not to be truly elementary, and when the number and variety of its constituents multiply (as has now happened with quarks), physicists typically entertain the hypothesis that *our* matter is not composed of any *really* elementary particles.

Suppose there is a part of the universe (possibly this one) in which matter *is* infinitely divisible. In that part of the universe there need be no upper bound on the amount of information storable in a given finite space. So my machine could perhaps exist, its tapes stored in a volume the size of, e.g., a human head. Indeed, one can imagine that where matter is infinitely divisible, there are creatures of all sizes, including creatures the size of electrons who agree to do the recording for us if we agree to wipe out their enemies by putting the lumps on which the enemies live in one of our particle accelerators.

Further, even if the story of the last paragraph is not nomologically possible, still it is not clear that the *kind* of nomological impossibility it possesses is relevant to my objection to the neo-Turing Test conception of intelligence. For if the neo-Turing Test conception of intelligence is an empirical "background" principle or law, it is a background principle or law of human *cognitive psychology*, not of *physics*. But a situation can contravene laws of physics without contravening laws of human psychology. For example, in a logically possible world in which gravity obeyed an inverse cube law instead of an inverse square law, our laws of *physics* would be different, but our laws of *psychology* might not be.

Now if my machine contravenes laws of nature, these laws are presumably laws of physics, not laws of psychology. For the question of how much information can be stored in a given space and how fast information can be transferred from place to place depends on such physical factors as the divisibility of matter and the speed of light. Even if the electron-sized creatures just described contravene laws of physics, still they need not contravene laws of human psychology. That is, humans (with their psychological laws intact) could coexist with the little creatures.[26]

[26] It may be objected that since brute force information processing methods

But if my machine does not contravene laws of human psychology—if it exists in a possible world in which the laws of human psychology are the same as they are here—then the neo-Turing Test conception of intelligence is false in a world where the laws of human psychology are the same as they are here. So the neo-Turing Test conception of intelligence cannot *be* one of the laws of human psychology.

In sum, the neo-Turing Test conception of intelligence can be construed either as some sort of conceptual truth or as a kind of psychological law. And it is false on both construals.

One final point: various sorts of modifications may make a variant of my machine nomologically possible in a much more straightforward sense. First, we could limit the vocabulary of the Turing Test to Basic English. Basic English has a vocabulary of only 850 words, as opposed to the hundreds of thousands of words in English, and it is claimed that Basic English is adequate for normal conversation, and for expression of a wide range of ideas. Second, the calculation made above was based on the string-searching version of the machine. The tree-searching version described earlier, however, avoids enormous amounts of duplication of parts of strings, and is no more intelligent.

More importantly, the machine as I have described it is designed to perform *perfectly* (barring breakdown); but perfect performance is far better than one could expect from any human, even ignoring strokes, heart attacks, and other forms of human "breakdown." Humans whose mental processes are functioning normally often misread sentences, or get confused; worse, any normal person engaged in a long Turing Test would soon get bored, and his attention would wander. Further, many loquacious souls would blather on from the very beginning, occasionally apologizing for not listening to the interlocutor. Many people would respond more by way of free association to the

---

are far more effective in the world in which matter is infinitely divisible than in ours, the laws of thought in that world *do* differ from the laws of thought in ours. But this objection begs the question, since if the string-searching machine I described cannot think in *any* world (as I would argue), the nomological difference which makes it possible is a difference in laws which affect the *simulation* of thought, not a difference in laws of thought.

interlocutor's remarks than by grasping their sense and giving a considered reply. Some people might devote nearly every remark to complaints about the unpleasantness of these interminable Turing Tests. If one sets one's sights on making a machine that does only as well in the Turing Test as *most* people would do, one might try a hybrid machine, containing a relatively small number of trees plus a bag of tricks of the sort used in Weizenbaum's program.

Perhaps many tricks can be found to reduce the memory load without making the machine any smarter. Of course, no matter how sophisticated the memory-reduction tricks we find, they merely postpone, but do not avoid the inevitable combinatorial explosion. For the whole point of the machine is to substitute memory for intelligence. So for a Turing Test of *some* length, perhaps a machine of the general type that I have described will be so large that making it any larger will cause collapse into a black hole. My point is that technical ingenuity being what it is, the point at which nomological impossibility sets in may be beyond what is required to simulate human conversational abilities.

Objection 7. The fault of the Turing Test as you describe it is one of experimental design, not experimental concept. The trouble is that *your* Turing Test has a *fixed length.* The programmers must know the length in order to program the machine. In an *adequate* version of the Turing Test, the duration of any occasion of testing would be decided in some random manner. In short, the trouble with your criticism is that you've set up a straw man.

Reply. It is certainly true that my machine's capacity to pass Turing Tests depends on there being some upper bound to the length of the tests. But the same is true of *people.* Even if we allow, say, twelve hours between question and answer to give people time to eat and sleep, still, people eventually *die.* Few humans could pass a Turing Test that lasted ninety years, and no humans could pass a Turing Test that lasted five hundred years. You can (if you like) characterize intelligence as the capacity to pass a Turing Test of arbitrary length, but since *humans do not have this capacity*, your characterization will not be a

necessary condition of intelligence, and even if it were a sufficient condition of intelligence (which I very much doubt—see below) a sufficient condition of intelligence that *humans do not satisfy* would be of little interest to proponents of views like the neo-Turing Test conception.

Even if medical advances remove any upper bound on the human life span, still people will die by accident. There is a nonzero probability that, in the course of normal thermal motion, the molecules in the two halves of one's body will move in opposite directions, tearing one in half. Indeed, the probability of escaping such accidental death literally *forever* is zero. Consider the "half-life" of people in a world in which death is put off as long as is physically possible. (The half-life for people, as for radioactive atoms, is the median life span, the time it takes for half to pass away.) Machines of my sort could be programmed to last for that half-life and (assuming they are no more susceptible to accidental destruction than people) their median life span would be as long as that of the median person.

Objection 7a. Let me try another tack. Cognitive psychologists and linguists often claim that cognitive mechanisms of one sort or another have "infinite capacities." For example, Chomsky says that our mechanisms for understanding language have the capacity to understand sentences of any length. An upper bound on the length of sentences people can understand in practice is a matter of interferences due to distraction, boredom, going mad, memory limitations, and many other phenomena, including, of course, death. This point is often put by saying that under the appropriate idealization (i.e., ignoring "interfering" phenomena of the sort mentioned) we have the capacity to understand sentences of any length. Now here is my point: under the same sort of idealization, we presumably have the capacity to pass a Turing Test of any length. But your string-searcher does *not* have this capacity, even under the appropriate idealization.

Reply. You seem to think you have objected to my claim, but really you have *capitulated* to it. I cheerfully concede that there is an idealization under which we probably have an "infinite" capacity that my machine lacks. But in order to describe this

idealization, you will have to indulge in a kind of theorizing about cognitive mechanisms that would be unacceptable to a behaviorist.

Consider the kind of reformulation of the neo-Turing Test conception of intelligence suggested by the idealization objection; it would be something like: "intelligence = the possession of language-processing mechanisms such that, were they provided with unlimited memory, and were they directed by motivational mechanisms that assigned at least a moderately high preference value to responding sensibly, and were they 'insulated' from 'stop' signals from emotion centers, and so forth, then the language-processing mechanisms would have the capacity to respond sensibly indefinitely." Notice that in order to state such a doctrine, one has to distinguish among various mental components and mechanisms. As an aside, it is worth noting that these distinctions have substantive empirical presuppositions. For example, memory might be inextricably bound up with language-processing mechanisms so as to make nonsense of talk of supplying the processing mechanisms with unlimited memory. The main point, however, is that in order to state such an "idealization" version of the neo-Turing Test conception one has to invoke mentalistic notions that no behaviorist could accept.

Objection 7b. I believe I can make my point without using mentalistic notions by idealizing away simply from nonaccidental causes of death. In replying to Objection 7, you said (correctly) that if medical advances removed an upper bound on human life, still the median string-searching machine could do as well as the median person. However, note that if nonaccidental causes of death were removed, every *individual* human would have no upper bound on how long *he* could go on in a Turing Test. By contrast, any individual string-searching machine must by its very nature have some upper bound on its ability to go on.

Reply. What determines how long we can go on in a Turing Test is not just how long we live, but the nature of our cognitive mechanisms and their interactions with other mental mechanisms. Suppose, for example, that we have no mechanisms for "erasing" information stored in long term memory. (Whether

this is so is not known.) If we can't "erase," then when our finite memories are "used up," normal conversational capacity will cease.

If the behaviorist identifies intelligence with the capacity to go on indefinitely in a Turing Test, idealizing away only from nonaccidental death, then people may turn out not to be intelligent in his sense. Further, even if people do turn out to satisfy such a condition, it can't be regarded as *necessary* for intelligence. Beings that go senile within two hundred years because they lack "erase" mechanisms can nonetheless be intelligent before they go senile.

Of course, the behaviorist could avoid this difficulty by further idealizing, explicitly mentioning erase mechanisms in his definition of intelligence. But that would land him back in the mentalistic swamp described in the last reply.

It is worth adding that even if we do have "erase" mechanisms, and even if nonaccidental causes of death were eliminated, still we would have *finite* memories. A variant of my string-searcher could perhaps exploit the finiteness of our memories so as to do as well as a person in an indefinitely long Turing Test. Suppose, for example, that human memory cannot record more than two hundred years of conversation. Then one of my string-searchers could perhaps be turned into a *loop-searcher* that could go on indefinitely. Instead of "linear" strings of conversation, it would contain circular strings whose ends rejoin the beginnings after, say, one thousand years of conversation. The construction of such loops would take much more inventiveness than the construction of ordinary strings. Even if it could be done, such a machine would seem intolerably repetitious to a being whose memory capacity far exceeded ours, but human conversation would seem equally repetitious to such a being.

Here is one final kind of rejoinder to the "unbounded Turing Test" objection. Consider a variant of my machine in which the programmers simply continue on and on, adding to the strings. When they need new tape, they reuse tape that has already been passed by.[27] Note that it is *logically* possible for the

[27] This machine would get ever larger unless the programmers were allowed to abandon strings which had been rendered useless by the course of the conversation. (In the tree-searching version, this would amount to pruning by-passed branches.)

everextending strings to come into existence by themselves—without the programmers (see note 21). Thus not even the capacity to go on indefinitely in a Turing Test is *logically* sufficient for intelligence.

Continuing on this theme, consider the infinitely divisible matter mentioned in the reply to Objection 6. It is logically and perhaps nomologically possible for a *man-sized* string-searching machine to contain creatures of everdecreasing size who work away making the tapes longer and longer without bound. Of course, neither of the two machines just mentioned has a *fixed* program, but since the programmers never see the stimuli, it is still the *machines* and not the programmers that are doing the responding. Contrast these machines with the infamous "machine" of long ago that contained a midget hidden inside it who listened to the questions and produced the answers.

Objection 8. You remarked earlier that the neo-Turing Test conception of intelligence is widespread in artificial intelligence circles. Still, your machine cannot be taken as refuting any AI (artificial intelligence) point of view, because as Newell and Simon point out, in the AI view, "the task of intelligence . . . is to avert the ever-present threat of the exponential explosion of search."[28] (In exponential explosion of search, adding one step to the task requires, e.g., 10 times the computational resources, adding two steps requires 10^2 (= 100) times the computational resources, adding three steps requires 10^3 (= 1000) times the computational resources, etc.) So it would be reasonable for AIers to amend their version of the neo-Turing conception of intelligence 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.*[29]

[28] Alan Newell and Herbert Simon, "Computer Science as Empirical Inquiry: Symbols and Search," in *Communications of the Association for Computing Machinery*, 19 (1979), 123.

[29] I am indebted to Dan Dennett for forcefully making this objection in his role as respondent to an earlier version of this paper in the University of Cincinnati Philosophy of Psychology Conference in 1978. Dennett tells me that he advocates the neo-Turing Test conception as amended above.

Reply. Let me begin by noting that for a proponent of the neo-Turing Test conception of intelligence to move to the *amended* neo-Turing Test conception is to capitulate to the psychologism that I have been defending. The *amended* neo-Turing Test conception attempts to avoid the problem I have posed by placing a *condition on the internal processing* (albeit a minimal one), viz., that it not be explosive. So the amended neo-Turing Test conception *does* characterize intelligence partly with respect to its internal etiology, and hence the amended neo-Turing Test conception is psychologistic.

While the amended neo-Turing Test conception is an improvement over the original neo-Turing Test conception in this one respect (it appeals to internal processing), it suffers from a variety of defects. One difficulty arises because there is an ambiguity in phrases such as "averts the exponential explosion of search." Such phrases can be understood as equivalent to "*avoids* exponential explosion altogether" (i.e., uses methods that do not require computational resources that go up exponentially with the "length" of the task) or, alternatively, as "*postpones* exponential explosion long enough" (i.e., *does* use methods that require computational resources that go up exponentially with the "length" of the task, but the "length" of the task is short enough that the required resources are in fact available). If it is postponing that is meant, my counterexample may well be untouched by the new proposal, because as I pointed out earlier, my machine or a variant on it may postpone combinatorial explosion long enough to pass a reasonable Turing Test.

On the other hand, if it is *avoiding* combinatorial explosion altogether that is meant, then the amended neo-Turing Test conception may brand *us* as unintelligent. For it is certainly possible that *our* information processing mechanisms—like those of many AI systems—are ones that succeed not because they avoid combinatorial explosion altogether, but only because they *postpone* combinatorial explosion long enough for practical purposes.

In sum, the amended neo-Turing Test conception is faced with a dilemma. If it is postponing combinatorial explosion that is meant, my machine may count as intelligent. If it is avoiding combinatorial explosion altogether that is meant, *we* (or other

intelligent organisms) may not count as intelligent.

Further, the proposed amendment to the neo-Turing Test conception is an entirely ad hoc addition. The trouble with such ad hoc exclusion of counterexamples is that one can never be sure whether someone will come up with another type of counterexample which will require another ad hoc maneuver.

I shall now back up this point by sketching a set of devices that have sensible input-output relations, but arguably are not intelligent.

Imagine a computer which simulates your responses to stimuli by computing *the trajectories* of all the elementary particles in your body. This machine starts with a specification of the positions, velocities, and charges (I assume Newtonian mechanics for convenience) of all your particles at one moment, and computes the changes of state of your body as a function of these initial conditions and energy impinging on your sensory mechanisms. Of course, what is especially relevant for the Turing Test is the effect of light from your teletype monitor on your typing fingers. Now though this takes some discussion, I opine that a machine that computes your elementary particle trajectories in this way is not intelligent, though it could control a robot which has the capacity to behave exactly as you would in any situation. It behaves as you do when you are doing philosophy, but *it* is not doing philosophy; rather, what it is doing is computing elementary particle trajectories so as to mimic your doing philosophy.

Perhaps what I have described is not nomologically possible. Indeed, it may be that even if God told us the positions and velocities of all the particles in your body, no computer could compute the complex interactions, even assuming Newtonian mechanics. However, notice one respect in which this machine may be superior to the one this paper has been mainly concerned with: namely, if it can simulate something for an hour, it may be able to simulate it for a year or a decade with the same apparatus. For continuing the simulation would be simply a matter of solving the same equations over and over again. For a wide variety of types of equations, solving the same equations over and over will involve no exponential explosion of search. If there is no exponential expansion of search here, the ad hoc

condition added in the objection is eluded, and we are left with the issues about nomological possibility that we discussed in Objection 6.

The idea of the machine just sketched could be applied in another machine which is closer to nomological possibility, namely one that simulates your *neurophysiology* instead of your elementary particle physics. That is, this machine would contain a representation of some adequate neurological theory (say, of the distant future) and a specification of the current states of all your neurons. It would simulate you by computing the changes of state of your neurons. Still more likely to be nomologically possible would be a machine which, in an analogous manner, simulates your *psychology*. That is, it contains a representation of some adequate psychological theory (of the distant future) and a specification of the current states of your psychological mechanisms. It simulates you by computing the changes of state of those mechanisms given their initial states and sensory inputs. Again, if there is no exponential expansion of search, the modification introduced in the objection gains nothing.

I said that these three devices are *arguably* unintelligent, but since I have little space to give any such arguments, this part of my case will have to remain incomplete. I will briefly sketch part of one argument.

Consider a device that simulates you by using a theory of your psychological processes. It is a robot that looks and acts as you would in any stimulus situation. Instead of a brain it has a computer equipped with a description of your psychological mechanisms. You receive a certain input, cogitate about it, and emit a certain output. If your robot doppelganger receives that input, a transducer converts the input into a description of the input. The computer uses its description of your cognitive mechanisms to deduce the product of your cogitations; it then transmits a description of your output to a mechanism that causes the robot body to execute the output. It is hardly obvious that the robot's process of manipulation of descriptions of your cogitation is *itself* cogitation. It is still less obvious that the robot's manipulation of descriptions of your experiential and emotional processes are themselves experiential and emotional processes.

To massage your intuitions about this a bit, substitute for the description-manipulating computer in your doppelganger's head a very small *intelligent person* who speaks only Chinese, and who possesses a manual (in Chinese) describing your psychological mechanisms. You get the input "Who is your favorite philosopher?" You cogitate a bit and reply "Heraclitus." Your robot doppelganger on the other hand contains a mechanism that transforms the question into a description of its sound; then the little man deduces that you would emit the noise "Heraclitus," and he causes the robot's voice box to emit that noise. The robot could simulate your input-output relations (and in a sense, simulate your internal processing, too) even though the person inside the robot understands nothing of the question or the answer. It seems that the robot simulates your thinking your thoughts without itself thinking those thoughts. Returning to the case where the robot has a description-manipulating computer instead of a description-manipulating person inside it, why should we suppose that the robot has or contains any thought processes at all?[30]

---

[30] Much more needs to be said to turn this remark into a serious argument. Intuitions about homunculi-headed creatures are too easily manipulable to stand on their own. For example, I once argued against functionalism by describing a robot that is functionally equivalent to a person, but is controlled by an "external brain" consisting of an army of people, each doing the job of a "square" in a machine table that describes a person. William Lycan objected ("Form, Function, and Feel," op. cit.) that the intuition that the aforementioned creature lacked mentality could be made to go away by imagining yourself reduced to the size of a molecule, and standing inside a person's sensory cortex. Seeing the molecules bounce about, it might seem absurd to you that what you were watching was a series of events that constituted or was crucial to some being's experience. Similarly, Lycan suggests, the intuition that my homunculi-heads lack qualia is an illusion produced by missing the forest for the trees, that is, by focusing on "the hectic activities of the little men, . . . seeing the homunculi-head as if through a microscope rather than as a whole macroscopic person." (David Rosenthal made the same objection in correspondence with me.)

While I think that the Lycan-Rosenthal point does genuinely alter one's intuitions, it can be avoided by considering a variant of the original example in which a *single* homunculus does the whole job, his *attention to column* S_i of a machine table posted in his compartment playing precisely the causal role required for the robot he controls to *have* S_i. (See "Are Absent Qualia Impossible?" op. cit., for a somewhat more detailed description of this case.) No "forest for the trees" illusion can be at work here. Nonetheless, the Lycan-

The string-searching machine with which this paper has been mainly concerned showed that behavior is intelligent only if it is *not* the product of a certain sort of information processing. Appealing to the Martian example described at the beginning of the paper, I cautioned against jumping to the conclusion that there is any positive characterization of the type of information processing underlying all intelligent behavior (except that it have at least a minimal degree of "richness"). However, what was said in connection with the Martian and string-searching examples left it open that though there is no single natural kind of information processing underlying all intelligent behavior, still there might be a kind of processing common to all *un*intelligent entities that nonetheless pass the Turing Test (viz., very simple processes operating over enormous memories). What this last machine suggests, however, is that it is also doubtful that there will be any interesting type of information processing common to such unintelligent devices.[31]

Massachusetts Institute of Technology

Rosenthal point does illustrate the manipulability of intuitions, and the danger of appealing to intuition without examining the source of the intuition. The role of most of the early objections and replies in this paper was to locate the source of our intuitions about the stupidity of the string-searching machine in its extremely simple information processing.

Another difficulty with the description-manipulator example is that it may seem that such an example could be used to show that *no symbol manipulation theory of thought processes* (such as those popular in cognitive psychology and artificial intelligence) *could be correct*, since one could always imagine a being in which the symbol-manipulating mechanisms are replaced by homunculi. (John Searle uses an example of the same sort as mine to make such a case in "Minds, Brains and Programs," forthcoming in *The Behavioral and Brain Sciences*. See my reply in the same issue.) While I cannot defend it here, I would claim that some symbol-manipulating homunculi-heads *are* intelligent, and that what justifies us in regarding some symbol-manipulating homunculus-heads (such as the one just described in the text) as unintelligent is that the causal relations among their states do not mirror the causal relations among our mental states.

[31] Previous versions of this paper were read at a number of universities and meetings, beginning with the 1977 meeting of the Association for Symbolic Logic. I am indebted to the following persons for comments on previous drafts: Sylvain Bromberger, Noam Chomsky, Jerry Fodor, Paul Horwich, Jerry Katz, Israel Krakowski, Robert Kirk, Philip Kitcher, David Lewis, Hugh Lacey, William Lycan, Charles Marks, Dan Osherson, Georges Rey, Sydney Shoemaker, George Smith, Judy Thomson, Richard Warner, and Scott Weinstein.