4

Skinner Skinned

B. F. Skinner has recently retired, after a long and distinguished career at Harvard, and for better or for worse it appears that the school of psychology he founded, Skinnerian behaviorism, is simultaneously retiring from the academic limelight. Skinner's army of enemies would like to believe, no doubt, that his doctrines are succumbing at last to their barrage of criticism and invective, but of course science doesn't behave like that, and the reasons for the decline in influence of behaviorism are at best only indirectly tied to the many attempts at its "refutation". We could soften the blow for Skinner, perhaps, by putting the unwelcome message in terms he favors: psychologists just don't find behaviorism very *reinforcing* these days. Skinner might think this was unfair, but if he demanded *reasons*, if he asked his critics to *justify* their refusal to follow his lead, he would have to violate his own doctrines and methods. Those of us who are not Skinnerians, on the other hand, can without inconsistency plumb the inner thought processes, reasons, motives, decisions and beliefs of both Skinner and his critics, and try to extract from them an analysis of what is wrong with Skinnerian behaviorism and why.

This is not an easy task, in large measure because of a spiralling escalation of vituperation between Skinner and his critics. Skinner began as a naive and achingly philistine social thinker, so the first rounds of *humanist* criticism of his position were contemptuous, and largely conducted in ignorance of Skinner's technical work or the background of theories against which it was developed. Skinner, recognizing this, did not conceal his contempt in turn for his arrogant and ignorant humanist opponents, and so it has continued, with both sides willfully misreading and misattributing, secure in the knowledge that

the other side is vastly underestimating its opponent. Somewhat surprisingly, Skinner's scientific critics have often fallen into similarly unedifying ruts.

Although counting myself among Skinner's opponents, I want to try to avoid the familiar brawl and do something diagnostic. I want to show *how* Skinner goes astray, through a series of all too common slight errors. He misapplies some perfectly good principles (principles, by the way, that his critics have often failed to recognize); he misdescribes crucial distinctions by lumping them all together; and he lets wishful thinking cloud his vision—a familiar enough failure. In particular, I want to show the falsehood of what I take to be Skinner's central philosophical claim, on which all the others rest, and which he apparently derives from his vision of psychology. The claim is that *behavioral science proves that people are not free, dignified, morally responsible agents.* It is this claim that secures what few links there are between Skinner's science and his politics. I want to show how Skinner arrives at this mistaken claim, and show how tempting in fact the path is. I would like to proceed by setting out with as much care as I can the steps of Skinner's argument for the claim, but that is impossible, since Skinner does not present arguments—at least, not wittingly. He has an ill-concealed disdain for arguments, a bias he feeds by supposing that brute facts will sweep away the most sophisticated arguments, and that the brute facts are on his side. His impatience with arguments does not, of course, prevent him from relying on arguments, it just prevents him from seeing that he is doing this—and it prevents him from seeing that his brute facts of behavior are not facts at all, but depend on an interpretation of the data which in turn depends on an an argument, which, finally, is fallacious. To get this phantom—but utterly central—argument out in the open will take a bit of reconstruction.

The first step in Skinner's argument is to characterize his enemy, "mentalism". He has a strong gut intuition that the *traditional* way of talking about and explaining human behavior—in "mentalistic" terms of a person's beliefs, desires, ideas, hopes, fears, feelings, emotions—is somehow utterly disqualified. This way of talking, he believes, is disqualified in the sense that not only is it not science as it stands; it could not be turned into science or used in science; it is inimical to science, would *have* to be in conflict with *any* genuine science of human behavior. Now the first thing one must come to understand is this antipathy of Skinner's for all things "mentalistic". Once one understands the antipathy, it is easy enough to see the boundaries of Skinner's enemy territory.

Skinner gives so many different reasons for disqualifying mentalism that we may be sure he has failed to hit the nail on the head—but he does get close to an important truth, and we can help him to get closer. Being a frugal Yankee, Skinner is reluctant to part with *any* reason, however unconvincing, for being against mentalism, but he does disassociate himself from some of the traditional arguments of behaviorists and other anti-mentalists at least to the extent of calling them relatively unimportant. For instance, perhaps the most ancient and familiar worry about mentalism is the suspicion that

(1) mental things must be made of *non-physical* stuff

thus raising the familiar and apparently fatal problems of Cartesian interactionism. Skinner presents this worry,[1] only to downplay it,[2] but when all else fails, he is happy to lean on it.[3] More explicitly, Skinner rejects the common behaviorist claim that it is

(2) the *privacy* of the mental

in contrast to the public objectivity of the data of behavior that makes the mental so abhorrent to science. "It would be foolish to deny the existence of that private world, but it is also foolish to assert that because it is private it is of a different nature from the world outside."[4] This concession to privacy is not all that it appears, however, for his concept of privacy is not the usual one encountered in the literature. Skinner does not even consider the possibility that one's mental life might be *in principle* private, *non-contingently* inaccessible. That is, he supposes without argument that the only sort of privacy envisaged is the sort that could someday be dispelled by poking around in the brain, and since "the skin is not that important as a boundary",[5] what it hides is nothing science will not be able to handle when the time comes. So Skinner suggests he will *not* object to the privacy of mental events, since their privacy would be no obstacle to science. At the same time Skinner often seeks to discredit explanations that appeal to some inner thing "we cannot see", which seems a contradiction.[6] For if we read these as objections to what we cannot *in principle* see, to what is necessarily unobservable, then he must after all be appealing tacitly to a form of the privacy objection. But perhaps we should read these disparagements of appeals to what we cannot see merely as disparagements of appeals to what we cannot *now* see, but whose existence we are *inferring.* Skinner often inveighs against appealing to

(3) events whose occurrence "can only be inferred".[7]

Chomsky takes this to be Skinner's prime objection against mentalistic psychology,[8] but Skinner elsewhere is happy to note that "Science often talks about things it cannot see or measure"[9] so it cannot be that simple. It is not that all inferred entities or events are taboo, for

Skinner himself on occasion explicitly infers the existence of such events; it must be a particular sort of inferred events. In particular,

(4) *internal* events

are decried, for they "have the effect of diverting attention from the external environment".[10] But if "the skin is not that important as a boundary", what can be wrong with internal events as such? No doubt Skinner finds *some* cause for suspicion in the mere internality of some processes; nothing else could explain his persistent ostrich-attitude towards physiological psychology.[11] But in his better moments he sees that there is nothing intrinsically wrong with inferring the existence of internal mediating events and processes—after all, he admits that some day physiology will describe the inner mechanisms that account for the relations between stimuli and responses, and he could hardly deny that in the meantime such inferences may illuminate the physiological investigations.[12] It must be only when the internal mediators are of a certain sort that they are anathema. But what sort? Why, the "occult", "prescientific", "fictional" sort, the "*mental* way station" sort,[13] but these characterizations beg the question. So the first four reasons Skinner cites are all inconclusive or contradicted by Skinner himself. If there is something wrong with mentalistic talk, it is not necessarily because mentalism is dualism, that mentalism posits nonphysical things, and it is not *just* that it involves internal, inferred, unobservable things, for he says or implies that there is nothing wrong with these features by themselves. If we are to go any further in characterizing Skinner's enemy we must read between the lines.[14]

In several places Skinner hints that what is bothering him is the *ease* with which mentalistic explanations can be concocted.[15] One *invents* whatever mental events one needs to "explain" the behavior in question. One falls back on the "miracle-working mind", which, just because it *is* miraculous, "explains nothing at all".[16] Now this is an ancient and honorable objection vividly characterized by Molière as the *virtus dormitiva*. The learned "doctor" in *Le Malade Imaginaire*, on being asked to explain what it was in the opium that put people to sleep, cites its *virtus dormitiva* or sleep-producing power. Leibniz similarly lampooned those who forged

> expressly occult qualities or faculties which they imagined to be like little demons or goblins capable of producing unceremoniously that which is demanded, just as if watches marked the hours by a certain horodeictic faculty without having need of wheels, or as if mills crushed grains by a fractive faculty without needing any thing resembling millstones.[17]

By seeming to offer an explanation, Skinner says, inventions of this sort "bring curiosity to an end". Now there can be no doubt that convicting a theory of relying on a *virtus dormitiva* is fatal to that theory, but getting the conviction is not always a simple matter—it often has been, though, in Twentieth Century psychology, and this may make Skinner complacent. Theories abounded in the early days of behaviorism which posited curiosity drives, the reduction of which explained why rats in mazes were curious; untapped reservoirs of aggressiveness to explain why animals were aggressive; and invisible, internal punishments and rewards that were postulated solely to account for the fact that unpunished, unrewarded animals sometimes refrained from or persisted in forms of behavior. But mentalistic explanations do not *seem* to cite a *virtus dormitiva*. For instance, explaining Tom's presence on the uptown bus by citing his desire to go to Macy's and his belief that Macy's is uptown does not look like citing a *virtus dormitiva*: it is not as empty and question-begging as citing a special uptown-bus-affinity in him would be. Yet I think it is clear that Skinner does think that all mentalistic explanation is infected with the *virtus dormitiva*.[18] This is interesting, for it means that *mentalistic* explanations are on a par for Skinner with a lot of bad *behavioristic* theorizing, but since he offers no discernible defense of this claim, and since I think the claim is ultimately indefensible (as I hope to make clear shortly), I think we must look elsewhere for Skinner's best reason for being against mentalism.

There is a special case of the *virtus dormitiva*, in fact alluded to in the Leibniz passage I quoted, which is the key to Skinner's objection: sometimes the thing the desperate theoretician postulates takes the form of a little man in the machine, a *homunculus*, a demon or goblin as Leibniz says. Skinner often alludes to this fellow. "The function of the inner man is to provide an explanation which will not be explained in turn."[19] In fact, Skinner identifies this little man with the notion of an autonomous, free and dignified moral agent: he says we must abolish "the autonomous man—the inner man, the homunculus, the possessing demon, the man defended by the literature of freedom and dignity".[20] This is a typical case of Skinner's exasperating habit of running together into a single undifferentiated lump a number of distinct factors that are related. Here the concept of a moral agent is identified with the concept of a little man in the brain, which in turn is identified with the demons of yore. Skinner, then, sees superstition and demonology every time a claim is made on behalf of moral responsibility, and every time a theory seems to be utilizing a homunculus. It all looks the same to him: bad. Moreover, he lumps *this* pernicious bit of

superstition (the moral-autonomous-homunculus-goblin) with all the lesser suspicions we have been examining; it turns out that "mental" means "internal" means "inferred" means "unobservable" means "private" means "*virtus dormitiva*" means "demons" means "superstition". Psychologists who study physiology (and hence look at *internal* things), or talk of *inferred* drives, or use mentalistic terms like "belief" are all a sorry lot for Skinner, scarcely distinguishable from folk who believe in witches, or, perish the thought, in the freedom and dignity of man. Skinner brands them all with what we might call guilt by free association. For instance, in *Beyond Freedom and Dignity*, after all Skinner's claims to disassociate himself from the lesser objections to mentalism, on p. 200 he lets all the sheep back into the fold:

> Science does not dehumanize man; it de-homunculizes him . . . Only by *dispossessing* him can we turn to the *real* causes of human behavior. Only then can we turn from the *inferred* to the observed, from the miraculous to the natural, from the *inaccessible* to the manipulable. (*my italics*)[21]

But I was saying that hidden in this pile of dubious and inconsequential objections to mentalism is something important and true. What is it? It is that Skinner sees—or almost sees—that there is a special way that questions can be begged in psychology, and this way is *akin to* introducing a homunculus. Since psychology's task is to account for the intelligence or rationality of men and animals, it cannot fulfill its task if anywhere along the line it *presupposes* intelligence or rationality. Now introducing a homunculus does just that, as Skinner recognizes explicitly in "Behaviorism at Fifty":

> . . . the little man . . . was recently the hero of a television program called "Gateways to the Mind" . . . The viewer learned, from animated cartoons, that when a man's finger is pricked, electrical impulses resembling flashes of lightning run up the afferent nerves and appear on a television screen in the brain. The little man wakes up, sees the flashing screen, reaches out, and pulls the lever . . . More flashes of lightning go down the nerves to the muscles, which then contract, as the finger is pulled away from the threatening stimulus. *The behavior of the homunculus was, of course, not explained.* An explanation would presumably require another film. And it, in turn, another. (*my italics*)[22]

This "explanation" of our ability to respond to pin-pricks depends on the intelligence or rationality of the little man looking at the TV screen in the brain—and what does *his* intelligence depend on? Skinner sees

clearly that introducing an unanalyzed homunculus is a dead end for psychology, and what he sees dimly is that a homunculus is hidden in effect in your explanation *whenever you use a certain vocabulary*, just because the use of that vocabulary, like the explicit introduction of a homunculus, presupposes intelligence or rationality. For instance, if I say that Tom is taking the uptown bus because he *wants* to go to Macy's and *believes* Macy's is uptown, my explanation of Tom's action *presupposes* Tom's intelligence, because if Tom weren't intelligent enough to put two and two together, as we say, he might fail to see that taking the uptown bus was a way of getting to Macy's. My explanation has a suppressed further premise: expanded it should read: Tom believes Macy's is uptown, and Tom wants to go to Macy's, so *since Tom is rational* Tom wants to go uptown, etc. Since I am relying on Tom's rationality to give me an explanation, it can hardly be an explanation of what makes Tom rational, even in part.

Whenever an explanation invokes the terms "want", "believe", "perceive", "think", "fear"—in short the "mentalistic" terms Skinner abhors—it must presuppose in some measure and fashion the rationality or intelligence of the entity being described.[23] My favorite example of this is the chess-playing computer. There are now computer programs that can play a respectable game of chess. If you want to predict or explain the moves the computer makes you can do it mechanistically (either by talking about the opening and closing of logic gates, etc., or at a more fundamental physical level by talking about the effects of the electrical energy moving through the computer) or you can say, "If the computer *wants* to capture my bishop and *believes* I wouldn't trade my queen for his knight, then the computer will move his pawn forward one space," or something like that. We need not take seriously the claim that the computer *really* has beliefs and desires in order to use this way of reasoning. Such reasoning about the computer's "reasoning" may in fact enable you to predict the computer's behavior quite well (if the computer is well-programmed), and in a sense such reasoning can even explain the computer's behavior—we might say: "Oh, now I understand why the computer didn't move its rook."—but in another sense it doesn't explain the computer's behavior at all. What is awesome and baffling about a chess-playing computer is how a mere mechanical thing could be made to be so "smart". Suppose you were to ask the designer, "How did the computer 'figure out' that it should move its knight?" and he replied: "Simple; it recognized that its opponent couldn't counterattack without losing a rook." This would be highly unsatisfactory to us, for the question is, how was he able to make a computer that *recognized* anything in the first place? So long

as our explanation still has "mentalistic" words like "recognize" and "figure out" and "want" and "believe" in it, it will presuppose the very set of capacities—whatever the capacities are that go to make up intelligence—it ought to be accounting for. And notice: this defect in the explanation need have nothing to do with postulating any non-physical, inner, private, inferred, unobservable events or processes, because it need not postulate any processes or events at all. The computer designer may know exactly what events are or are not going on inside the computer, or for that matter on its highly visible output device: in choosing to answer by talking of the computer's *reasons* for making the move it did, he is not asserting that there are any extra, strange, hidden processes going on; he is simply explaining the *rationale* of the program without telling us how it's done. Skinner comes very close to seeing this. He says:

> Nor can we escape. . . . by breaking the little man into pieces and dealing with his wishes, cognitions, motives, and so on, bit by bit. The objection is not that those things are mental but that they offer no real explanation and stand in the way of a more effective analysis.[24]

The upshot of this long and winding path through Skinner's various objections to mentalism is this: if we ignore the inconsistencies, clear away the red herrings, focus some of Skinner's vaguer comments, and put a few words in his mouth, he comes up identifying the enemy as a certain class of terms—the "mentalistic" terms in his jargon—which when used in psychological theories "offer no real explanation" because using them is something like supposing there is a little man in the brain. Skinner never says the use of these terms presupposes rationality, but it does. Skinner also never gives us an exhaustive list of the mentalistic terms, or a definition of the class, but once again we can help him out. These terms, the use of which presupposes the rationality of the entity under investigation, are what philosophers call the *intentional idioms.*[25] They can be distinguished from other terms by several peculiarities of their logic, which is a more manageable way of distinguishing them than Skinner's.[26] Thus, spruced up, Skinner's position becomes the following: *don't use intentional idioms in psychology.*

Spruced-up Skinner is not alone in being opposed to intentional idioms in psychology. His Harvard colleague, Quine, has been most explicit on the topic.[27] One might suppose their congruence on this issue came out of discussion or collaboration, but Skinner is so apparently oblivious of Quine's arguments against intentional psychology, and so diffuse in his own objections to "mentalism" that this is most unlikely. For Quine's objections to intentional idioms have never had anything to do with their presupposing rationality or offering no explanation; rather he has argued that intentional idioms are to be foresworn because, as Chisholm argues, we cannot translate sentences containing intentional idioms into sentences lacking them.[28] Sentences containing intentional idioms refuse to "reduce" to the sentences of the physical sciences, so we must learn to do without them; Skinner on the other hand is blithely confident that such translations are possible,[29] and indeed *Beyond Freedom and Dignity* consists in large measure of samples of Skinner's translations.[30]

If Skinner never avails himself of the Chisholm-Quine untranslatability argument, and never makes explicit the presupposition of rationality argument, he does nevertheless muddy the water with a few other inconclusive objections. Intentional explanations tend to be "unfinished", he says, in that an action is explained, for instance, by reference to an opinion, without the existence of the opinion being explained in turn. But explaining an explosion by citing a spark is similarly incomplete, and since Skinner admits that both the former and the latter explanation could be completed, this is hardly a telling objection.[31] He also suggests that intentional explanations are not predictive, which is manifestly false. (See Chapters 1 and 15 of this volume.) Knowing that Tom wants to go to Macy's and believes the uptown bus will take him there, my prediction that he will take the uptown bus is, while not foolproof, highly reliable. Skinner sometimes hints that intentional explanations are only *vaguely* predictive, but this does not distinguish them from his own explanations until we are given some parameters by which to measure vagueness, which for human behavior are not forthcoming.

So let us put words in Skinner's mouth, and follow the phantom argument to its conclusion. We can, then, "agree" with Skinner when we read him between the lines to be asserting that no satisfactory psychological theory can *rest* on any use of intentional idioms, for their use presupposes rationality, which is the very thing psychology is supposed to explain. So if there is progress in psychology, it will inevitably be, as Skinner suggests, in the direction of eliminating ultimate appeals to beliefs, desires, and other intentional items from our explanations. So far so good. But now Skinner appears to make an important misstep, for he seems to draw the further conclusion that *intentional idioms therefore have no legitimate place in any psychological theory.* But this has not been shown at all. There is no reason why

intentional terms cannot be used provisionally in the effort to map out the functions of the behavior control system of men and animals, just so long as a way is found eventually to "cash them out" by designing a mechanism to function as specified (see Chapters 1, 5 and 7). For example, we may not now be able to describe mechanically how to build a "belief store" for a man or animal, but if we specify how such a belief store must function, we can use the notion in a perfectly scientific way pending completion of its mechanical or physiological analysis. Mendelian genetics, for instance, thrived as a science for years with nothing more to feed on than the concept of a gene, a whatever-it-turns-out-to-be that functions as a transmitter of a heritable trait. All that is required by sound canons of scientific practice is that we not suppose or claim that we have reached an end to explanation in citing such a thing. Skinner, or rather phantom-Skinner, is wrong, then, to think it follows from the fact that psychology cannot make any *final appeal* to intentional items, that there can be no place for intentional idioms in psychology.

It is this misstep that leads Skinner into his most pervasive confusion. We have already seen that Skinner, unlike Quine, thinks that translation of intentional into non-intentional terms is possible. But if so, why can't intentional explanations, in virtue of these bonds of translation, find a place in psychology? Skinner vacillates between saying they can and they can't, often within the space of a few pages.

> Beliefs, preferences, perceptions, needs, purposes, and opinions are possessions of autonomous man which are *said to* change when we change minds. What *is* changed in each case is a probability of action. (*my italics*)[32]

How are we to interpret this? As meaning that we change probabilities, *not* beliefs, or as meaning that changing beliefs *is just* changing probabilities of action? Skinner's very next sentence strongly suggests the latter:

> A person's belief that the floor will hold him as he walks across it depends upon his past experience.

but a few sentences later he hedges this by putting "belief" in scare-quotes:

> We build "belief" when we increase the probability of action by reinforcing behavior.

Does this passage mean that it is *all right* to talk of building belief, so long as we understand it as increasing action probabilities, or that it is

wrong to talk that way since *all* we are doing is increasing action probabilities?[33] On the next page he takes the hard line:

> We change the relative strengths of responses by differential reinforcement of alternative courses of action; we do *not* change something called a preference. We change the probability of an act by changing a condition of deprivation or aversive stimulation; we do *not* change a need. We reinforce behavior in particular ways; we do *not* give a person a purpose or an intention. (*my italics*)

This vacillation is typical of Skinner. The exclusivity expressed in the last quotation is rampant in *Beyond Freedom and Dignity*: "Our age is not suffering from anxiety but from the accidents, crimes, . . ." (p. 14) Young people refuse to get jobs "not because they feel alienated but because of defective social environments . . ." (p. 15) A man "makes a distinction not through some mental act of perception but because of prior contingencies". (p. 187) (See also, pp. 26, 30, 157, 101, 189, 190, 204) Yet the contrary claim that these intentional terms can all be translated, and hence, presumably, can be used to make true statements in psychology, is just about as widespread. We have just seen what may be Skinner's definition of "believe"; "want" is defined on p. 37, and "intend" on p. 72, and p. 108. Intentional idioms occur by the dozens in crucial roles in all of Skinner's books, and Skinner explicitly justifies or excuses this practice in several places. For instance, in *Beyond Freedom and Dignity* (p. 24), he says, "No doubt many of the mentalistic expressions imbedded in the English language cannot be as rigorously translated as 'sunrise', but acceptable translations are not out of reach." In *About Behaviorism* (p. 17) he says, "Many of these expressions I 'translate into behavior'. I do so while acknowledging that *traduttori traditori*—translators are traitors—and that there are perhaps no exact behavioral equivalents . . ." But the context shows that Skinner thinks he only loses the flavor—the connotations—not the predictive or inferential power or referential accuracy of the terms.

It is unfathomable how Skinner can be so sloppy on this score, for reflection should reveal to him, as it will to us, that this vacillation is over an absolutely central point in his argument.[34] For surely Skinner is right in seeing that the validity of our conceptual scheme of moral agents having dignity, freedom and responsibility stands or falls on the question: can men ever be *truly* said to have beliefs, desires, intentions? If they can, there is at least some hope of retaining a notion of the

dignity of man; if they cannot, if men never can be said truly to want or believe, then surely they never can be said truly to act responsibly, or to have a conception of justice, or know the difference between right and wrong. So Skinner's whole case comes down to the question: can intentional explanations (citing beliefs, desires, etc.) on the one hand, and proper, ultimate, scientific explanations on the other hand, *co-exist*? Can they ever *both* be true, or would the truth of a scientific explanation always *exclude* the other?

In spite of his vacillation in print, it is clear that Skinner must come down in favor of the exclusive view, if his argument is to work. Certainly the majority of his remarks favor this view, and in fact it becomes quite explicit on p. 101 of *Beyond Freedom and Dignity* where Skinner distinguishes the "pre-scientific" (i.e., intentional) view of a person's behavior from the scientific view, and goes on to say, "Neither view can be proved, but it is in the nature of scientific inquiry that the evidence should shift in favor of the second." Here we see Skinner going beyond the correct intuition that it is in the nature of scientific inquiry that ultimate appeals to intentional idioms must disappear as progress is made, to the bolder view that as this occurs intentional explanations will be rendered false, not reduced or translated into other terms.

I argue at length in Chapter 12 that intentional and mechanistic or scientific explanations *can* co-exist, and have given here an example supposed to confirm this: we know that there is a purely mechanistic explanation of the chess playing computer, and yet it is *not false* to say that the computer *figures out* or *recognizes* the best move, or that it *concludes* that its opponent cannot make a certain move, any more than it is false to say that a computer *adds* or *multiplies*. There has often been confusion on this score. It used to be popular to say, "A computer can't really think, or course; all it can do is add, subtract, multiply and divide." That leaves the way open to saying, "A computer can't really multiply, of course; all it can do is add numbers together very, very fast," and that must lead to the admission: "A computer cannot really add numbers, of course; all it can do is control the opening and closing of hundreds of tiny switches," which leads to: "A computer can't really control its switches, of course; it's simply at the mercy of the electrical currents pulsing through it." What this chain of claims adds up to "prove", obviously, is that computers are really pretty dull lumps of stuff—they can't do anything interesting at all. They can't really guide rockets to the moon, or make out paychecks, or beat human beings at chess, but of course they can do all that and more. What the computer programmer can do

if we give him the chance is not *explain away* the illusion that the computer is doing these things, but *explain how* the computer truly is doing these things.

Skinner fails to see the distinction between explaining and explaining away. In this regard he is succumbing to the same confusion as those who suppose that since color can be explained in terms of the properties of atoms which are not colored, nothing is colored. Imagine the Skinner-style exclusion claim: "The American flag is *not* red, white and blue, but rather a collection of colorless atoms." Since Skinner fails to make this distinction, he is led to the exclusive view, the view that true scientific explanations will exclude true intentional explanations, and typically, though he asserts this, he offers no arguments for it. Once again, however, with a little extrapolation we can see what perfectly good insights led Skinner to this error.

There are times when a mechanistic explanation obviously does exclude an intentional explanation. Wooldridge gives us a vivid example:

> When the time comes for egg laying the wasp *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into her burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralyzed cricket, which has not decayed, having been kept in the wasp equivalent of deep freeze. To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavor of logic and thoughtfulness—until more details are examined. For example, the wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If, while the wasp is inside making her preliminary inspection the cricket is moved a few inches away, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again the wasp will move the cricket up to the threshold and re-enter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion, this procedure was repeated forty times, always with the same result.[35]

In this case what we took at first to be a bit of intelligent behavior is

unmasked. When we see how simple, rigid and mechanical it is, we realize that we were attributing too much to the wasp. Now Skinner's experimental life has been devoted to unmasking, over and over again, the behavior of pigeons and other lower animals. In "Behaviorism at Fifty" he gives an example almost as graphic as our wasp. Students watch a pigeon being conditioned to turn in a clockwise circle, and Skinner asks them to describe what they have observed. They all talk of the pigeon *expecting, hoping* for food, *feeling* this, *observing* that, and Skinner points out with glee that they have observed nothing of the kind; he has a simpler, more mechanical explanation of what has happened, and it *falsifies* the students' unfounded *inferences.* Since in this case explanation is unmasking or explaining away, it always is.[36] Today pigeons, tomorrow the world. What Skinner fails to see is that it is not the fact that he has an explanation[37] that unmasks the pretender after intelligence, but rather that his explanation is so simple (see Chapter 12 of this volume). If Skinner had said to his students, "Aha! You think the pigeon is so smart, but here's how it learned to do its trick," and proceeded to inundate them with hundreds of pages of detailed explanation of highly complex inner mechanisms, their response would no doubt be that yes, the pigeon did seem, on his explanation, to be pretty smart.

The fact that it is the simplicity of explanations that can render elaborate intentional explanations false is completely lost to Skinner for a very good reason: the only *well-formulated, testable* explanations Skinner and his colleagues have so far come up with have been, perforce, relatively simple, and deal with the relatively simple behavior controls of relatively simple animals. Since all the explanations he has so far come up with have been of the unmasking variety (pigeons, it turns out, do not have either freedom or dignity), Skinner might be forgiven for supposing that all explanations in psychology, including all explanations of human behavior, must be similarly unmasking.

It might, of course, turn out to be the case that all human behavior could be unmasked, that all signs of human cleverness are as illusory as the wasp's performance, but in spite of all Skinner's claims of triumph in explaining human behavior, his own testimony reveals this to be wishful thinking. Even if we were to leave unchallenged all the claims of operant conditioning of human beings in experimental situations,[38] there remain areas of human behavior that prove completely intractable to Skinner's mode of analysis. Not surprisingly, these are the areas of deliberate, intentional action. The persistently recalcitrant features

of human behavior for the Skinnerians can be grouped under the headings of novelty and generality. The Skinnerian must explain all behavior by citing the subject's past history of similar stimuli and responses, so when someone behaves in a novel manner, there is a problem. Pigeons do not exhibit very interesting novel behavior, but human beings do. Suppose, to borrow one of Skinner's examples, I am held up and asked for my wallet.[39] This has never happened to me before, so the correct response cannot have been "reinforced" for me, yet I do the smart thing: I hand over my wallet. Why? The Skinnerian must claim that this is not truly novel behavior at all, but an instance of a *general sort* of behavior which has been previously conditioned. But what sort is it? Not only have I not been trained to hand over my wallet to men with guns, I have not been trained to empty my pockets for women with bombs, nor to turn over my possessions to armed entities. None of these things has ever happened to me before. I may never have been threatened before at all. Or more plausibly, it may well be that most often when I have been threatened in the past, the "reinforced" response was to *apologize* to the threatener for something I'd said. Obviously, though, when told, "Your money or your life!" I don't respond by saying, "I'm sorry. I take it all back." It is perfectly clear that what experience has taught me is that if I *want* to save my skin, and *believe* I am being threatened, I should do what I *believe* my threatener *wants* me to do. But of course Skinner cannot permit this intentional formulation at all, for in ascribing wants and beliefs it would presuppose my rationality. He must insist that the "threat stimuli" I now encounter (and these are not defined) are similar in some crucial but undescribed respect to some stimuli encountered in my past which were followed by responses of some sort similar to the one I now make, where the past responses were reinforced somehow by their consequences. But see what Skinner is doing here. He is positing an external *virtus dormitiva.* He has no record of any earlier experiences of this sort, but *infers* their existence, and moreover *endows* them with an automatically theory-satisfying quality: these postulated earlier experiences are claimed to resemble-in-whatever-is-the-crucial-respect the situation they must resemble for the Skinnerian explanation to work. Why do I hand over my wallet? Because I must have had in the past some experiences that reinforced wallet-handing-over behavior in circumstances like this.

When Skinner predicts pigeon behavior he makes use of his knowledge of their reinforcement history, but when he predicts human behavior, he does not. This can be vividly seen if we consider once

again the chess-playing computer. Suppose we set up a contest between Skinner and an intentionalist to see who could make the best predictions of the computer's moves. Skinner would proceed by keeping a careful cumulative record of every move the computer ever made, keeping track of each move's consequences, to see which moves were "reinforced" by their consequences. Would he have a chance of making good predictions? Mathematically, it can be shown that there is no guarantee that he would get anywhere with this method unless he knew the *internal* starting state of the computer; the computer's past biography of moves is not enough.[40] But for the sake of argument we can suppose that the importance of the initial state would recede as the computer made more and more moves (not a universally plausible supposition), so that Skinner could get closer and closer to good predictions in spite of his ignorance of this crucial variable. Skinner's predictions would take this form: there is a high probability that the computer will move queen to king's bishop-4 because when stimulated by similar (not necessarily identical) board positions in the past, the computer has been reinforced for making similar (not necessarily identical) moves. There is obviously much that is problematical about such formulations—e.g., what are the shared features of the similar board positions and similar moves?—but let us suppose Skinner succeeds, after years of cumulative recording, in arriving at good predictions. This would not be as flashy and easy a method as the intentionalist's, who would simply ask himself at each point in the game: "Now if I were the computer, knowing what I know and wanting what I want, which move would I believe to have the best consequences?" but Skinner could comfort himself by recalling Russell's phrase, and claiming that his opponent's method had all the advantages of theft over honest toil.

But suppose we complicate the picture. Suppose we wrote some chess-playing programs that could "learn" as they played, and improve as a result of "experience"—by the relatively simple expedient of adjusting weightings in their evaluation formulae for positions as a function of their "track record". Now suppose we set two different chess-playing computers to playing a series of games against each other, but do not provide for the recording of the games. We turn them on at night, and in the morning discover two very much improved chess-playing computers (one of them, probably, would have established its mastery over the other—something we couldn't discover until we watched a few games). It must be possible to determine mathematically what these evolved programs are now (if we know exactly what

program each had to begin with, and that there is no randomizing element in either program, and that no uncorrected malfunctions occurred during the night, and if we know exactly how many games were played in the time available)—or we could, in principle, take the computers apart, figuring out what their programs were in that way. But for practical purposes both of these methods are ruled out. Can we still make predictions of their behavior? Of course. The intentionalist can predict their behavior just as well (no better) than he can predict the behavior of a novel human opponent, a stranger in town. He would assume his opponent had some intelligence, and hence would expect him to make the most intelligent moves available. But Skinner would have to claim ignorance; the fact that the biography of the computers would be lost would mean that Skinner would not be able to use his method. He would say that too much conditioning of which he had no record had intervened overnight. But he *could* do this: he could make the same predictions (roughly, depending on his ability at chess) as the intentionalist, and *on the same grounds*, namely that it was the best move he could see, and then "deduce" the fact that the computer during the night *must have* been "reinforced" for making moves "similar in some respect" to the one he is now predicting. Here it would be crystal clear that Skinner would have no grounds for such a hypothesis except that his theory required it, and no way of being specific about the "similarity" of the overnight experiences.[41]

I am suggesting that once Skinner turns from pigeons to people, his proffered "explanations" of human behavior are no better than this. If Skinner complains that mentalistic explanations are too easy, since we always know exactly what mental events to postulate to "explain" the behavior, the same can be said of all the explanation sketches of complex human behavior in Skinner's books. They offer not a shred of confirmation that Skinner's basic mode of explanation—in terms of reinforcement of operants—will prove fruitful in accounting for human behavior. It is hard to be sure, but Skinner even seems to realize this. He says at one point, "The instances of behavior cited in what follows are not offered as 'proof' of the interpretation", but he goes right on to say, "The proof is to be found in the basic analysis." But insofar as the "basic analysis" proves anything, it proves that people are not like pigeons, that Skinner's unmasking explanations will not be forthcoming. Certainly if we discovered that people only handed over their wallets to robbers after being conditioned to do this, and, moreover, continued to hand over their wallets after the robber had shown his gun was empty, or when the robber was flanked by policemen, we

would have to admit that Skinner had unmasked the pretenders; human beings would be little better than pigeons or wasps, and we would have to agree that we had no freedom and dignity.

Skinner's increasing reliance, however, on a *virtus dormitiva* to "explain" complex human behavior is a measure of the difference between pigeons and persons, and hence is a measure of the distance between Skinner's premises and his conclusions. When Skinner speculates about the past history of reinforcement in a person in order to explain some current behavior, he is saying, in effect, "I don't know which of many possible equivalent series of events occurred, but one of them did, and that explains the occurrence of this behavior now." But what is the equivalence class Skinner is pointing to in every case? What do the wide variety of possible stimulus histories have in common? Skinner can't tell us in his vocabulary, but it is easy enough to say: the stimulus histories that belong to the equivalence class have in common the fact that they *had the effect of teaching the person that p*, of storing certain information. In the end Skinner is playing the same game with his speculations as the cognitivist who speculates about internal representations of information. Skinner is simply relying on a more cumbersome vocabulary.

Skinner has failed to show that psychology without mentalism is either possible or—in his own work—actual, and so he has failed to explode the myths of freedom and dignity. Since that explosion was to have been his first shot in a proposed social revolution, its misfiring saves us the work of taking seriously his alternately dreary and terrifying proposals for improving the world.

5

Why the Law of Effect Will Not Go Away

The poet Paul Valéry said: "It takes two to invent anything." He was not referring to collaborative partnerships between people but to a bifurcation in the individual inventor. "The one", he says, "makes up combinations; the other one chooses, recognizes what he wishes and what is important to him in the mass of the things which the former has imparted to him. What we call genius is much less the work of the first one than the readiness of the second one to grasp the value of what has been laid before him and to choose it."[1] This is a plausible claim. Why? Is it true? If it is, what kind of truth is it? An empirical generalization for which there is wide scale confirmation? Or a "conceptual truth" derivable from our concept of invention? Or something else?

Herbert Simon, in *The Sciences of the Artificial*, makes a related claim: "human problem solving, from the most blundering to the most insightful, involves nothing more than varying mixtures of trial and error and selectivity."[2] This claim is also plausible, I think, but less so. Simon presents it as if it were the conclusion of an inductive investigation, but *that*, I think, is not plausible at all. An extensive survey of human problem solving may have driven home this thesis to Simon, but its claim to our assent comes from a different quarter.

I want to show that these claims owe their plausibility to the fact that they are implications of an abstract principle whose "necessity" (such as it is) consists in this: we can know independently of empirical research in psychology that any adequate and complete psychological theory must exploit some version or other of the principle. The most familiar version of the principle I have in mind is the derided darling of the behaviorists: the Law of Effect. "The rough idea", Broadbent observes,[3] "that actions followed by reward are repeated, is one which