https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

AUTONOMY FROM THE VIEWPOINT OF TELEOLOGICAL BEHAVIORISM

By Howard Rachlin

I. Introduction

I will argue that the autonomy of a particular act of a particular person depends on the pattern of behavior in which it is embedded. I call this *conditional autonomy*. A person's act is conditionally autonomous or not, relative to other acts at other times. Consider an example of a person crossing the street. On the one hand, this act might not be done for its own sake, but may fit into some ongoing long-term behavioral pattern that is personally beneficial to the person crossing the street—such as regularly buying groceries in the supermarket (which happens to be across the street). On the other hand, crossing the street might be done simply for its own sake. If such an act were considered to be autonomous, regardless of its temporal context, its autonomy would be *unconditional*. However, I will argue that whereas conditional autonomy is a highly useful social concept, indeed a necessary concept, for any human society, unconditional autonomy is a useless concept that actually impedes our efforts to understand and explain human behavior.

Before making this argument I need to discuss the psychological viewpoint from which it comes, a viewpoint I call *teleological behaviorism*. In order to do this, I first need to distinguish between cognitive psychology and behavioral psychology (as I understand these terms), and then, to distinguish between Skinnerian behaviorism (the form of behaviorism that most people are familiar with) and teleological behaviorism.

II. Cognitive Psychology and Behavioral Psychology

Figure 1 shows, in a general way, how behavioral and cognitive theorists conceptualize choice behavior. The thick vertical line represents the boundary between the person and the world. To the right of the thick line is an example of a cognitive organization of behavior; to the left, a behavioral organization. The four heavy horizontal arrows represent the variables of both cognitive and behavioral psychologies—events in the world affecting the person and behavior of the person observable in the world. The upper arrow heading from the world into the person represents information—signals relevant to important events in the world. The upper arrow going to the left is verbal behavior—what the person

© 2003 Social Philosophy & Policy Foundation. Printed in the USA.

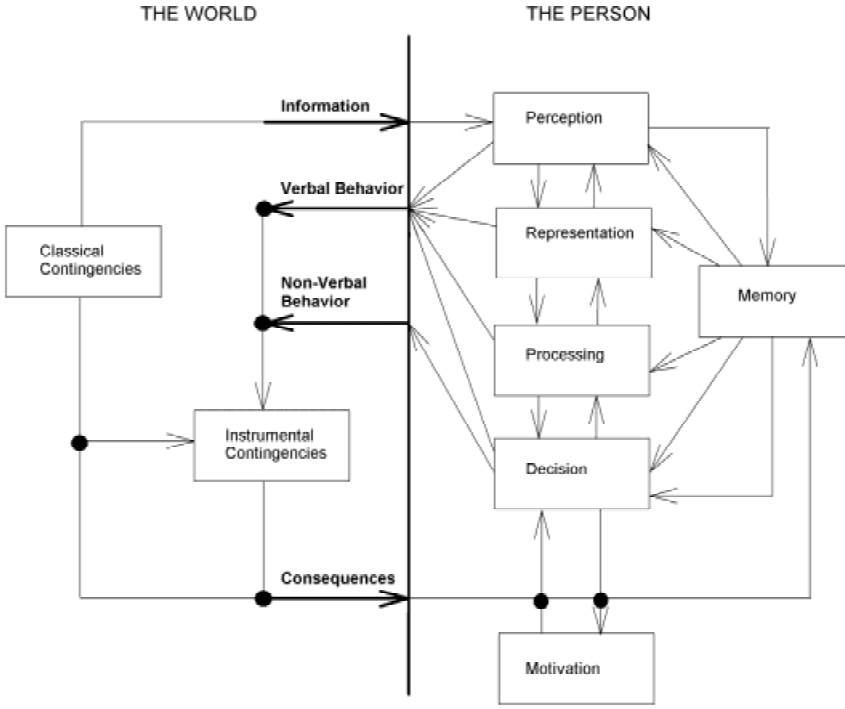


FIGURE 1. Cognitive and behavioral models

says and writes. The lower arrow going to the left is nonverbal behavior—the actual choices that a person makes. The lower arrow going to the right indicates consequences, or outcomes, of his or her choices.

The elements of cognitive theory (in very simplistic and abstract form, not intended to reflect any particular theory) are the internal mechanisms that mediate between the arrows going in and the arrows going out between the person and the world. (See the boxes and thin arrows to the right of the thick vertical line.) Information, in the form of instructions and experimental displays, is first perceived, then represented and processed, and finally a decision is made on the basis of it. All of these processes are influenced by memory. In addition, processes may be altered by motives coming up from below. The workings of some or all of these mechanisms are (in some cognitive theories) accessible to introspection and expressible by verbal behavior. When a decision is made, that decision is expressed in verbal or nonverbal behavior or some combination of the two. Finally, the outcome of the decision (consequences) is retained in memory and may affect future operation of any component of the decision mechanism.

Figure 1 is a simplistic model and does not indicate any of the quite complex and elaborate submechanisms (neural networks or parallel pro-

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

cessing) that may operate within or across the boxes in the figure. There is some dispute within cognitive psychology about which, if any, internal mechanisms contain submechanisms that work on logical principles, and where, if anywhere in the process, principles of optimality apply. Moreover, strictly speaking, a purely cognitive theory is silent about the physiological instantiation (the hardware) corresponding to the flow of information (the software) represented in Figure 1. The workings of the boxes labeled perception, representation, memory, and so forth, and their interaction as an information processing system, are what cognitive theories are all about.

What goes on to the right of the heavy vertical line of Figure 1 (that is, what goes on in the cognitive realm) is unquestionably interesting and important. No behaviorist can deny that. Nevertheless, the behaviorist is fundamentally interested, not in the principles underlying internal mechanisms, but in the principles underlying the behavior itself. The cognitivist and the behaviorist both want, ultimately, to *explain* current behavior—to answer the question, "Why is this person acting in this way?" Their explanations of current behavior, however, refer to different events. The cognitivist, in explaining current actions, refers to current events inside the organism; the behaviorist, in explaining current actions, refers to past actions and past environmental events, and to the relations over time between those events and actions. The behavioral equivalent to a cognitive mechanism is a contingency. But, unlike a cognitive mechanism, a contingency has no meaning at an instant of time. For example, a student's grade point average, or GPA, is presumably contingent on his or her study habits; but no particular GPA value can be traced to any particular act of studying. Study habits are definable only over extended time periods.

No behaviorist, not even a teleological behaviorist like me, denies that there are mechanisms inside the person mediating between the incoming and outgoing arrows of Figure 1. The organism is not empty. But, just as the cognitive psychologist may be interested fundamentally in the internal software and only secondarily, if at all, in the particular underlying hardware (the particular brain physiology), so the behaviorist is interested fundamentally in the relations in the world or environment between the four thick arrows of Figure 1, and only secondarily in the mechanisms (software or hardware) that mediate between the ingoing and outgoing arrows.

III. Skinnerian Behaviorism

Skinnerian analysis proceeds as follows. There is some behavior in which you are interested (say, addiction). First you look for a single stimulus in the environment that reliably causes that behavior. I think that most reasonable people will agree that there is no such single environ-

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

mental cause of addiction; if there were, we could cure addiction simply by eliminating its stimulus. Failing to find a stimulus, Skinnerian analysis bids you to look for a *reinforcer*, to ask, "What are the consequences of any given bit of behavior?"

Suppose that I receive a written invitation to speak at a conference. I prepare the talk, go to the conference, and deliver the talk. (Presumably this is voluntary behavior.) You could ask *how* I do it, tracing a chain of internal events from the typed words entering my eyes, engaging my cognitive system, and finally causing me to stand up and speak. Or you could ask, "What are the reinforcers (and punishers)?" This would involve an examination of the consequences of my behavior as they affect me—and these might include the chance to enjoy the attention of my audience for a while, or perhaps the chance to influence them, or to improve my research, thanks to feedback from audience members. You would have to look for similar behavior by me in the past and similar consequences. These, Skinner says, are the true causes of my present behavior. I may have better access to my own behavior and its consequences than the audience could have (because I am always there when I am behaving), and this is an advantage for me. However, my audience has an objective view that I do not have, and that is an advantage for them. (The behavior therapist aims to work with this advantage.)

What is wrong with this method of analyzing behavior? The problem is that, just as it is not possible, in almost any interesting case, to find a single external stimulus as the cause of behavior, it is likewise not possible to find an external reinforcer in many interesting cases. For example, although a member of my audience might easily discover reinforcers for my talk, he will not be able to find a reinforcer for any particular sentence. Yet, according to Skinner, every individual act (each sentence, each word) must have its individual, immediate reinforcer—either a primary reinforcer (like a pellet of food offered to a rat) or a secondary reinforcer (like the mere sound of the food hopper moving in a Skinner box). It is often impossible, though, to trace every individual act of a laboratory rat (let alone a human being) to its particular reinforcement.

This problem is apparent in my primary research area: self-control. On the one hand, it is easy to explain *failures* of self-control in terms of immediate reinforcement. If I have a second or third cocktail at a party, the reinforcers of my behavior—the cocktails themselves—are easy to spot. But what if the waiter tries to hand me a cocktail, and I refuse? What reinforces that act? What reinforces *success* in self-control? I refuse the cocktail and exactly nothing happens.[1] The usual tactic of the behavior

[1] One could say that not vomiting and not getting into embarrassing social situations are discrete reinforcers of a particular drink refusal. But if the absence of these particular aversive events is to explain a particular refusal, then their absence must be distinguished from the absence of any of the infinite number of other conceivable aversive events (experiencing electric shock, being boiled in oil, etc.) that are not contingent on drinking but are

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

analyst at this point is to postulate some sort of inner self-reinforcement—my satisfaction perhaps. When I refuse the third drink I might covertly pat myself on the back for resisting temptation. (I will discuss self-reinforcement in more detail later in this essay.) But to postulate such an internal mechanism is to engage in cognitive analysis rather than behavioral analysis (as I have defined it). Cognitive psychologists are much better at postulating internal mechanisms than behaviorists are—cognitive psychologists are trained for it—and the behaviorist begins to look overly simplistic. This is the problem with Skinnerian behaviorism, which is a problem that has been pointed out many times.[2]

IV. Teleological Behaviorism

When confronted with a clearly voluntary act (like refusing a cocktail) for which there is no apparent external reinforcement, teleological behaviorism looks, not into the organism, but into the past and future, for patterns of behavior.[3] According to teleological behaviorism, the question of why I am giving a talk cannot be broken up into the questions of why I am speaking this sentence, why I am speaking the next sentence, and so forth. Rather, teleological behaviorism examines why I am giving the talk as a whole—the *extrinsic* social reinforcers as well as the *intrinsic* value of the activity itself.[4] If I had to decide whether to utter each sentence individually, I would be tongue-tied, since there is no reinforcement for each sentence individually. Similarly, in the area of self-control, there is no reinforcement for each act of refusing a cocktail (to cite my earlier example) or, for that matter, for refusing an individual cigarette, hit of heroin, or so forth. It is sometimes said that the reinforcement for such acts is *delayed*. But if I refuse a single cocktail tonight I will not wake up three weeks from today, suddenly a healthier, happier person. A single act of self-control is not reinforced—*now or ever*. Reinforcement is to be found,

---

also generally absent after drink refusal. The most straightforward way to make this distinction is to point to the negative correlation over time between drink refusal and vomiting or getting into embarrassing social situations (i.e., the more drink refusal, the less vomiting there is). But, in Skinner's model, negative correlations cannot work by themselves. They must be reduced to specific reinforcers following specific responses. In order to effect this reduction, Skinnerian behaviorism has to postulate unobserved internal stimuli, responses, and reinforcers.

[2] William Baum, *Understanding Behaviorism: Science, Behavior, and Culture* (New York: Harper Collins, 1994); John Staddon, *The New Behaviorism: Mind, Mechanism, and Society* (Philadelphia, PA: Psychology Press, 2001).

[3] Howard Rachlin, *Behavior and Mind: The Roots of Modern Psychology* (New York: Oxford University Press, 1994); Howard Rachlin, "Teleological Behaviorism," *American Psychologist* 47 (1992): 1371–82.

[4] Intrinsic value is determined not by introspection or physiological or biological investigation, but by the position of the pattern on a value scale or in a utility function. Value and utility are determined, in turn, by behavioral observation. Eating, for example, is a highly valued act, not because it is pleasurable or stimulates a "pleasure center" in the brain, but because eating is frequently chosen over other acts.

not in individual acts, but in the value (or utility) of a pattern of acts repeated over time. This value may be extrinsic to the pattern—such as social approval—or intrinsic in the pattern itself, such as the enhanced enjoyment of food when eating is well regulated.[5]

To put this argument more formally, suppose that there is a long activity lasting T time units and a short activity lasting t time units where $T = Nt$. A problem of self-control arises when two conditions are satisfied:

$$\text{A: } V_T > NV_t, \text{ and}$$

$$\text{B: } V_t > V_{T/N},$$

where V stands for value. In other words, a problem of self-control arises when the value of a longer activity is greater than that of an equal duration of repetitions of a shorter activity and the value of the shorter activity is greater than that of an equal fraction of the longer activity.[6]

For example, suppose that you were driving from New York to California. You like both classical and popular music, so you take along ten classical compact discs (CDs) and ten pop CDs to play on your car's CD player during the trip. Let us assume that the order of your preference (illustrated in Figure 2) is as follows:

(1) listening to a whole one-hour-long symphony;
(2) listening to a three-minute-long popular song;
(3) listening to three minutes of the symphony.

The problem is, of course, that in order to listen to the whole symphony (first on your list) you must listen to the first three minutes of the symphony (last on your list). Hence, the immediate choice is always between (3) and (2) (the popular song). If you always do what you immediately prefer, you are likely to reach California without ever having heard the symphony. Of course, you could listen to just one popular song and then the symphony, but after listening to the song, you would be in the same position that you were in originally. The value of the symphony (as the gestalt psychologists never tire of pointing out) lies not in the sum of the values of its parts—even in the sum of the delay-discounted values of its parts; it lies in the symphony as a whole. You can see this by imagining how you would feel if a CD were defective and the last three minutes of a symphony were missing. You would not consider that you had 57/60th

[5] In dealing with individual human behavior, the behaviorist calculates reinforcement value in terms of consummatory behavior rather than in terms of the goods themselves being bought. For an individual, value (or utility for that matter) is primarily in eating the apple and only secondarily in having the apple.

[6] By an *activity* I mean an instance of a class of activities defined functionally (an *operant* in Skinner's terms). Thus, eating a large steak for dinner and eating a pint of ice cream that night while watching TV would be two instances of the same class of activity—overeating.

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

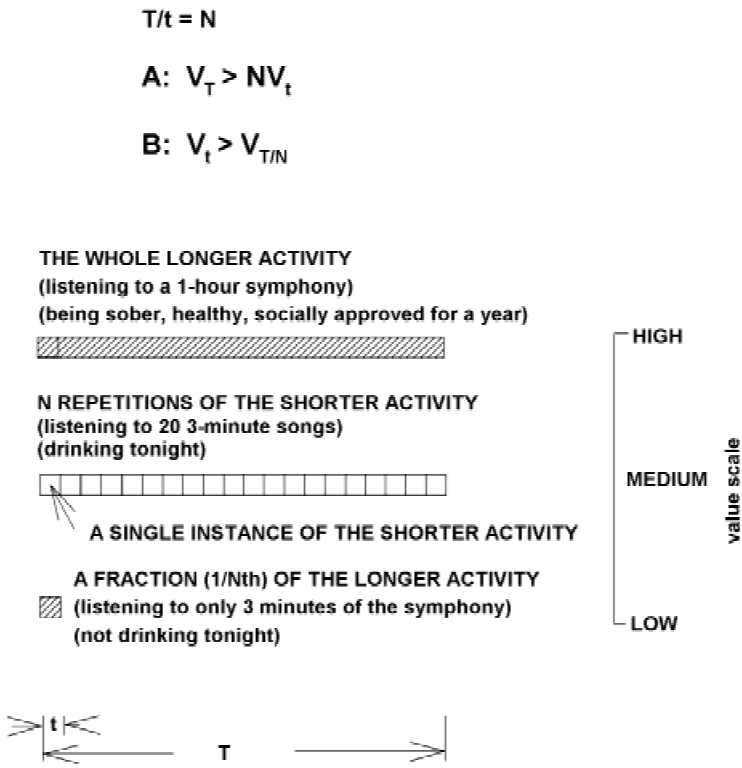


FIGURE 2. How inconsistency of valuation of long and short duration activities cause problems in self-control

worth of enjoyment from the CD. Most likely, the whole experience would be ruined, even though you had already listened to, and presumably enjoyed, fifty-seven minutes of the symphony.

The same considerations apply in most instances of human self-control—for example, the choice between studying and not studying, between reading a novel and watching a TV program, or between being sober and having too much to drink. The alcoholic does not choose to be an alcoholic. He prefers not to be one. His preference ordering is:

(1) not being an alcoholic;
(2) drinking tonight;
(3) not drinking tonight.

He just chooses to drink tonight, and tonight, and tonight—and ends up as an alcoholic without ever having chosen to be one.

The cognitive psychologist naturally sees self-control in terms of internal mechanisms; the teleological behaviorist sees self-control as more or

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

less valuable behavioral patterns, reinforced in different behavioral contexts *as whole patterns*. This difference between the behavioral and cognitive approaches is not just a semantic issue. It determines where you should look—into the person or into the person's past (the person's reinforcement history, in Skinner's terms)—to discover the causes of her behavior.

V. Negative and Positive Addictions

A negative addiction is the behavioral pattern that we are most accustomed to calling addictive—that of the alcoholic, or the heroin or cocaine addict, or the cigarette smoker.[7] Negative addictions come about not because they are chosen as such, but because of two prior properties. First, there exists an inverse proportionality between the rate of the addictive activity and the value of each addictive act. If the rate of the addictive activity is initially low, the value of each alcoholic drink, drug dose, or cigarette is high. But as the rate increases, tolerance develops, reducing the value of each alcoholic drink, drug dose, or cigarette. The alcoholic, drinking at a high rate, needs a quart of gin to achieve the effect that a single martini would have in a social drinker. Second, as high rates of addictive activity reduce the value of each addictive act, they also reduce the value of all alternatives to that act. That is, negative addictions reduce the quality of life. As alcoholism develops, for example (and the value of each individual drink drops), relationships with family, friends, coworkers, and the drinker's general health also deteriorate. If, as alcoholism progresses, the rate of deterioration of the value of alternative activities deteriorates as fast or faster than that of drinking itself, drinking will remain the most highly valued of all available activities, or even gain in relative value, as it loses in absolute value. As a character in Linda Yablonski's novel *Junk* says, "However good or bad you feel, heroin makes you feel better."[8] At low rates, choosing the addictive activity increases an already high overall value; at high rates, choosing the addictive activity keeps an already low overall value from decreasing further.[9]

Positive addictions work in an opposite manner to negative ones. As the rate of a positive addictive activity increases, the value of an individual episode increases as well; as the rate decreases, the value of an individual episode decreases as well. Such addictions require learning or practice to attain high value: examples include listening to classical music,

[7] The distinction between negative and positive addictions comes from George J. Stigler and Gary S. Becker, "De Gustibus Non Est Disputandum," *American Economic Review* 67 (1977): 76–90.

[8] Linda Yablonski, *The Story of Junk: A Novel* (New York: Ferrer, Straus, and Giroux, 1997), as quoted in Sue Halpern, "The Awful Truth," *New York Review of Books*, September 25, 1997, p. 13.

[9] Technically, as the rate increases the addictive activity changes from a positive reinforcer to a negative reinforcer.

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

exercising, doing crossword puzzles, or collecting stamps. Most importantly, *social* activity is a *positive* addiction. As the rate of social activity increases, the value of each positively addictive act increases. If the value of any alternative to the act does not increase faster than that of the act, the act's increased value will cause further increases in its rate, which will, in turn, increase its value still further, and so on, resulting in the positive addictions we observe.

In the case of social activity, a child's appropriate social behavior must be externally reinforced in order to be maintained.[10] That is, a child must learn to behave appropriately in a given society. As we grow older, the more we engage in social activity, the more social support we get from other people in return. Experiments in my laboratory with undergraduate subjects have shown that socially cooperative behavior depends on the probability that it will be reciprocated by others.[11] Thus, social activity may be extrinsically reinforced.

Along with offering *extrinsic* reinforcement, social activity, when it forms a coherent pattern, is *intrinsically* of high value. As in the previous examples of positive addictions, like doing crossword puzzles, listening to classical music, engaging in sports, etc., social activity beyond a certain rate is performed for its own sake (i.e., frequently chosen over other activities). Yet, the less we engage in social activities, the more we retire into our own shells, the more difficult it becomes to start social activity again, and the less intrinsically valuable social activity becomes. The same holds for other positive addictions. For example, I have been skiing four times in my life, each separated far enough from the other times that I had to learn all over again how to ski. I never graduated from the beginner's slope. Needless to say, I do not enjoy skiing.

Although positive addictions develop in a manner opposite to negative ones, there is evidence that positive addictions are substitutable for negative addictions. I have proposed a theory of addiction, called *relative addiction theory*, which says that the most fundamental way (although not necessarily the only way or even the most effective way) to get rid of a negative addiction is to substitute a positive addiction for it.[12] Effecting such substitutions, however, is not a simple matter, since all during the recovery the particular negatively addictive act will be higher in value than the particular positively addictive act: the alcoholic will always prefer the local bar to dinner with his family. What is required, as in most acts of self-control, is for choices to be made between patterns of acts extending over time rather than between particular acts.

[10] Mary D. S. Ainsworth and John Bowlby, "An Ethological Approach to Personality Development," *American Psychologist* 46 (1991): 333–41.

[11] Forest Baker and Howard Rachlin, "Probability of Reciprocation in Prisoner's Dilemma Games," *Journal of Behavioral Decision Making* 14 (2001): 51–67.

[12] Howard Rachlin, "Four Teleological Theories of Addiction," *Psychonomic Bulletin and Review* 4 (1997): 462–73.

As in the example, in the previous section, of the CD with the missing three minutes of music at the end, the value of the act at the beginning of the sequence is *provisional*. The value of the first fifty-seven minutes of listening to the hour-long symphony will depend on what happens during the last three minutes—even though those fifty-seven minutes are in the past and presumably have already been experienced. The same would be true for doing a crossword puzzle. My enjoyment of the puzzle depends strongly on whether I finish it. But my enjoyment or, very frequently, frustration, applies to the whole of the time I spend doing the puzzle.[13] My choice to listen to the symphony or do the puzzle, therefore, must be a choice to listen to the *whole* symphony or do the *whole* puzzle; the value of these activities derives from the patterns of the activities, not from the sums of the values of their parts.

VI. Unconditional Autonomy

Many of the philosophical difficulties with the concept of autonomy come from insisting that we be able to classify a given act as unconditionally autonomous or unconditionally not autonomous before we know what pattern the act is part of. The belief that we can classify particular acts unconditionally comes from a cognitive view of behavior. In a cognitive model, everything you would want to know about a particular act is given by the instantaneous state of the various internal mechanisms that form the efficient cause of the act. If the state of those mechanisms were known—either by inference from past and present behavior, or by introspection by the behaving person—the value, the degree of autonomy, the purpose, the meaning, and the relevant context of the act would also be known. But, from the viewpoint of teleological behaviorism, none of these qualities can be known absolutely—all are provisional.[14]

Imagine that you see a snippet of film that shows a man swinging a hammer. But what is he *actually* doing? Now consider the following alternative descriptions. He is:

(1) swinging a hammer
(2) hammering a nail
(3) joining one piece of wood to another
(4) building a floor
(5) building a house

[13] The test is whether, after discovering the missing three-minute final section of the CD, I play the CD again. (I would not play it again; if you would, then what if six or nine or twelve minutes were missing? At some point the whole CD must lose its value.)

[14] A probability is provisional in the same sense. A given probability can always change depending on future events. But, like a probability, as more and more behavior is observed the meaning of a particular act can be known better and better—by the observer as well as (or better than) by the actor.

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

(6) providing shelter for his family
(7) supporting his family
(8) being a good husband and father
(9) being a good person

All of these may be valid descriptions of his behavior. Based on the snippet of film you saw, all you can say is that he is swinging the hammer. He might have been swinging it at someone's head. But if you said, "He's just swinging the hammer," someone else who saw more of the film might, with justice, say, "Yes, he's swinging the hammer, but what he's *really* doing is hammering a nail." A third person who has seen even more of the film might then correct the second in a similar way: "Yes, he's hammering a nail, but what he's *really* doing is building a floor." We can envision such a process until we get to a godlike observer who has sat through a film of virtually this man's entire life and can make the final judgment: "He's being a good person." To paraphrase Ovid (*Metamorphoses*, III): No one must be counted good until his last funeral rites are paid.

The important point is that all of the descriptions, listed from (1) through (9), are descriptions of the man's behavior. As you go down the list, more and more context is incorporated into the description, but that context is always behavioral. The final godlike observer need not look into the man's heart as long as he has looked at the complete film of the man's life. Then the observer can say, "He is a good man," without fear of contradiction—regardless of the state of the man's heart. The validity of any of the above descriptions may be settled by moving the camera back or by showing more film, earlier and later.

This teleological and behavioristic way of assessing an activity is different from the usual way of doing so. Usually, as you go down a list of possible descriptions or explanations, you are supposed to be going deeper and deeper into the man himself—deeper and deeper into his internal intentions, his internal consciousness, his nervous system, or his soul (where the soul is considered to be some nonmaterial, internal entity). In a sense, the usual conception has to be true: if the man is building a house, some internal mechanism must be causing his movements. Correspondingly, if a sound system is playing Beethoven's Fifth Symphony, there must be some mechanism that causes what we hear—AM, FM, LP, CD, or tape. But the mechanism is not the thing we are describing. For a sound system to be playing Beethoven's Fifth Symphony, Beethoven's Fifth Symphony must be coming out of the system, regardless of the mechanism. Similarly, for a man to support his family, he must support his family, regardless of the mechanism that prompts him to behave.

You might say that there is a difference between the man who intentionally swings a hammer and the man who accidently swings a hammer—even though the two men are behaving alike. You might also say that the

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

difference resides inside them. But, again, while it is true that something inside of people must mediate behavioral differences, the difference between intention and accident *is* behavioral. The difference between a man purposely swinging a hammer and a man accidently swinging a hammer can be resolved, not by looking inside him, but by looking at more of his behavior. A man accidently swinging a hammer will not be hammering a nail or building a floor.

VII. Conditional Autonomy

A useful distinction between autonomous behavior and nonautonomous behavior can be based on the distinction between self-controlled and non-self-controlled (or impulsive) behavior. Such a distinction may help in attributing responsibility to people for their own actions. A person would then be deemed responsible for his actions to the degree that those actions were self-controlled. Attributing responsibility allows, in turn, the efficient construction of social contingencies of reward and punishment. But there would be no point in rewarding or punishing actions that cannot be controlled by reward and punishment. Therefore, to classify an act as autonomous, it must be controllable by its consequences.

Skinner called acts that are controllable by their consequences "operants." He distinguished operants from what he called "respondents," acts controlled by their antecedents. The antecedents to which Skinner referred were the classical environmental stimuli—the food powder injected into the mouth of a hungry dog causing salivation, the blow of a rubber hammer on the patellar tendon causing a knee jerk.[15]

Respondents were said to be *elicited*; operants were said to be *emitted*. A behavioral conception of unconditional autonomy, if such were possible, might rest on this distinction: emitted acts (acts controllable by their consequences) would be autonomous; elicited acts (acts controllable by their antecedents) would not be autonomous. The problem with this notion is that *any* act—even salivation, pupillary constriction, or heart rate—can be controlled, at least to some extent, by its consequences. Moreover, some acts, clearly controlled by their consequences (such as an alcoholic's drinking), would seem most usefully classified as nonautonomous.

The literature on addiction often refers to addiction as slavery. I believe that a useful definition of autonomy should take this commonsense notion into account. The addict is a slave, however, not because some inner

[15] Some critics of Skinner, including some philosophers, seem to believe that the postulation of such stimuli-response (S-R) connections was his entire goal. Their criticisms, if they are criticisms of the notion that psychology's goal is to track down every response to its immediately antecedent stimulus, should really be directed toward cognitive and physiological psychology. It is these psychologists who, if they cannot find an immediately antecedent stimulus for a response in the environment, undertake to search for one in the interior of the organism. And, if they cannot find one, they invent one and give it a name—such as willpower.

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

force is compelling each movement he makes. Reaching out for the glass, grasping it, lifting it to his mouth, are all physiologically voluntary acts, not reflexes. The addict is a slave because the reinforcement of his act is strong and immediate, however brief.

A person with a positive addiction, on the other hand, is not a slave. Once the long-term pattern of a positive addiction attains its high intrinsic value, each act that comprises that pattern is, I would say, an autonomous act. Each act has, by itself, no extrinsic reinforcement, immediate or delayed. Nor is it intrinsically reinforced; on the contrary, its immediate alternatives may be highly preferred to it. Its value is contingent on being embedded within a pattern of other acts. That is what makes it autonomous.

An unconditional autonomy, based merely on whether or not an act can be controlled by its consequences, would include virtually any individual act, including those acts society would want to classify as nonautonomous, and such an inclusive conception would therefore be useless. Realizing this, Skinner rejected the concept of autonomous actions entirely.

Is there a behavioral conception of autonomy that would make a useful social distinction between acts for which a person is responsible and acts for which a person is not responsible? To answer this question, we need to expand Skinner's operant/respondent classification from two to four kinds of acts: (1) *Respondents*, reflexes clearly attributable to an external stimulus; (2) *Emitted operants*, acts of high intrinsic value, like eating, that are done for their own sake; (3) *Reinforced operants*, acts, like paying for a loaf of bread, that are done not for their own sake, but for the sake of an extrinsic reinforcer; and (4) *Self-controlled acts*, acts, like pedaling on an exercise bicycle or refusing a drink, that are of low intrinsic value yet never extrinsically reinforced, which are nevertheless part of a pattern of acts of high intrinsic value (a healthy life or an active social life, for example). Acts in this last category are conditionally autonomous.

Aristotle drew a famous analogy between the pattern of an individual person's acts over time and the pattern of acts of a group of people. In *Posterior Analytics* (II, 19) he said that the pattern of acts (the universal) comes about through individual actions (the particulars) "like a rout in battle stopped by first one man making a stand and then another, until the original formation has been restored."[16] In a battle, one soldier alone knows that he cannot possibly survive the onslaught of a pursuing army. That soldier also knows that his action can potentially influence the actions of others and, therefore, establish a pattern of resistance. Similarly, in ordinary life, one night's abstinence from a bad habit would be futile against the pursuing army of temptations. It would be futile, that is, were it not for the connections in our behavior between one night and the

[16] Aristotle, *Posterior Analytics*, trans. E. M. Edghill, in Richard McKeon, ed., *The Basic Works of Aristotle* (New York: Random House, 1941).

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

next—were it not for our ability to organize our behavior into patterns extending over time (as a military formation extends over space). That is, some particular acts are not intrinsically valuable in themselves, but become valuable when massed into patterns (or habits).

I have argued that at least some of these patterns are valuable for society, as well as for the individual.[17] Bicycling to work rather than driving, for example, may be preferable in the long run for a person (in terms of health gained), as well as for society (in terms of pollution avoided), but it might not be preferred to driving in the short run (in terms of personal discomfort and energy expended). *Although particular acts of social cooperation may not be intrinsically valued by the particular person performing them, the pattern of acts, of which the individual acts are a part, may be intrinsically valuable to that person*. In other words, the problem of self-control and that of social control are intertwined. While patterns of behavior that are beneficial to society may also be intrinsically valuable for the individual person in the long run, the individual acts that comprise those patterns may not be intrinsically valuable and would not be chosen over their alternatives without extrinsic reinforcement (or punishment of their alternatives). From the viewpoint of teleological behaviorism, society's primary function, *vis-à-vis* its members, is to arrange contingencies of reinforcement and punishment so as to establish these beneficial patterns.

What then makes an act autonomous? A particular act is autonomous when it *would not* be chosen over its particular alternatives, but is part of an intrinsically valuable pattern of behavior that *would* be chosen over alternative patterns. The alcoholic, for example, might not refuse a particular drink but might at the same time prefer not to be an alcoholic. Thus, his refusal of the drink would be an autonomous act relative to his drinking the drink. How contingencies of reinforcement and punishment might best be arranged, so that people will pattern their behavior and not make decisions on a case-by-case basis, is an empirical question and the subject of my research on self-control.[18]

VIII. Internalization

The teleological behavioral approach to autonomy is completely external. No particular internal processes are presupposed. The degree to which an act is autonomous does not depend on the actor's introspection or feeling of freedom, nor on the action or lack of action of any of the mechanisms in the boxes on the right side of Figure 1.[19]

[17] Howard Rachlin, "Altruism and Selfishness," *Behavioral and Brain Sciences* (forthcoming).

[18] Howard Rachlin, *The Science of Self-Control* (Cambridge, MA: Harvard University Press, 2000).

[19] It depends to some extent on the *perception* of freedom, but the perception is that of the observer (or that of the actor as an observer of the consequences of his own actions).

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

Consider the sequence of events as a complex pattern is learned—a girl learning to play the piano, for example. At first, each particular key press must be learned individually and reinforced by the teacher, but soon chords and chord sequences become learning units. Then, a whole musical piece may be reinforced—by the teacher's approval, by the girl's parents, by the applause of her audience. Finally, perhaps, piano playing becomes intrinsically reinforced; that is, the girl may reach a stage at which playing the piano becomes valuable to her in itself. She plays the piano even when all extrinsic reinforcers are removed. Reaching this point (and she certainly might not reach it—even, or perhaps especially, if she becomes a professional musician) does not mean that reinforcement has become internalized. It does not mean that her playing is reinforced by internal applause, internal pats on her internal back, internal satisfaction, or internal stimulation of some "pleasure center" in her brain. It means that, when offered repeated choices between playing a whole sonata (with no extrinsic reinforcement) and other valued activities requiring equal time (say, watching a TV program or reading a magazine article), she frequently chooses to play the sonata.

From the viewpoint of teleological behaviorism, moreover, the intrinsic value of her piano playing refers not to what she *would do* now, at this particular point in time, but to what she *does* in the long run, including what she does in the future. Thus, the value of any pattern of acts is always provisional. Like the value of the first fifty-seven minutes of the symphony on the CD in the example presented previously, the value of the girl's piano playing depends on future events, as well as past events. The moral implication of this conception is that future behavior may have the power to redeem past behavior—by altering its context. Whether a professional pianist's present concert (followed by applause and money) is intrinsically valued by her will depend on her future, as well as her past, behavior (whether she continues to play the piano in her retirement, for example).

Most psychologists, including many who call themselves behaviorists, would not accept this behavioral view of reinforcement. They believe that the processes of reinforcement and punishment may be internalized. Positive internal events such as satisfaction are supposed to reinforce internal responses; negative internal events such as guilt are supposed to punish internal responses. The whole left part of Figure 1 is transposed into an internal representation within the right part. This internal mechanism is then held to be responsible for autonomous behavior, understood in terms of some form of self-regulation, self-governance, or self-direction.

The concept of internal reinforcement and punishment has both logical and empirical problems. As early as 1896, John Dewey argued that con-

However, the observer's perception in this case is his *overt* behavior of classifying acts as more or less autonomous.

cepts such as *stimulus*, *response*, and the then-recently postulated *reinforcement* had meaning only in terms of input to and output from the organism as a whole.[20] Inside the organism there are physiological events such as nerve firings, which could respond to prior nerve firings and stimulate other nerve firings, but there is no clear point (unless there is an homunculus) at which stimuli on one side could be divided from responses on the other.[21]

More recently, philosopher Rowland Stout has developed this critique, arguing against what he calls "the internal shift," that is, the process of taking concepts defined originally in terms of behavior of whole organisms over time, reifying them, and projecting the reified object inside the organism.[22]

Despite their logical problems, the concepts of self-reinforcement and self-punishment have survived and even thrived in psychology. As one would suspect, they are very difficult to test. Since self-reinforcement and self-punishment are supposed to occur wholly within the person, they are impossible to observe. By definition, satisfaction is a pleasant emotion and guilt is a painful one. But there is no evidence that these emotions can reinforce or punish behavior—reinforce the refusal of a drink or punish the taking of it, for example. An alcoholic is more likely to take another drink to get rid of the guilt of having taken the previous one. Like slapping yourself on the hand or hitting your head with the heel of your hand, or even kicking yourself, guilt may serve to enhance the feedback—to remind you that you did, indeed, have the drink. But guilt does not punish the taking of the drink, as an electric shock punishes a rat's pressing a lever in a Skinner box. If an alcoholic could press a button, *before* he had the drink, so as to deliver an electric shock *after* he had the drink, that would be an effective form of self-control. This is similar to the way Antabuse (a drug that makes drinking painful) works. But, just as alcoholics generally refuse to take Antabuse, the alcoholic would probably not press the button. Once the drink is taken, pressing the button would be futile. It would be no more effective as punishment than slapping yourself on the hand.

Some people say that they reinforce their own writing behavior, for example, by keeping a bowl of peanuts on the table and eating one after each page. "Self-reinforcement" of this kind may actually help a person to write. If so, how does it work? The problem for the theorist is to explain why the peanut is ever withheld. A reinforcer must be more valuable than

[20] John Dewey, "The Reflex-Arc Concept in Psychology," *Psychological Review* 3 (1896): 357–70.

[21] It would be possible to draw a line around the brain and then to speak of stimuli to the brain and responses of it, but all theories of internal reinforcement proposed thus far have reinforcement taking place *within* the brain.

[22] Rowland Stout, *Things That Happen Because They Should* (Oxford: Oxford University Press, 1996).

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

the reinforced act. If peanuts can reinforce writing a page, eating a peanut would have to be more valuable than writing a page; otherwise, why not eat a peanut after each paragraph, after each sentence, after each word, or, since it is preferred to writing, why not eat a peanut before writing or instead of writing? Writing may well be extrinsically reinforced—by fame, money, tenure. But can it be reinforced by peanut eating? If the other rewards were taken away, and if writing an article were not intrinsically reinforcing, could such an activity be reinforced by eating peanuts?

In an attempt to study how self-reinforcement actually works, a group of dieters were invited to "reinforce" their own weight loss by taking money from a dish placed next to the scale at weekly weigh-ins. Sure enough, the dieters who did this lost more weight than another group with no dish of money next to the scale. However, a third group, whose members actually put money into the dish proportional to the amount of weight lost—and, thus, if self-punishment were possible, would have been "punishing" themselves for weight lost—actually lost the most weight.[23] The true reinforcers in this experiment were social approval and health gain, not the few dollars taken from a dish. Taking the money served as vivid feedback—it underscored the behavior. Taking the money was a way of saying, "Yes, I did lose the weight and will get the reward." Putting money into the dish evidently served as even more vivid feedback. Thus, with all of its logical difficulties, the concept of self-reinforcement apparently labors under empirical difficulties, too.

IX. Objections to a Behavioral View of Autonomy

There is something admittedly unsatisfying about this wholly behavioral view of autonomy—as there is about behavioral explanations in general. However, I would like to emphasize that the usual objection to behaviorism—that two people may do the same thing, but do it for different reasons—does not apply to teleological behaviorism. Even though two people may be doing the same thing at this very moment, if they are doing it for different reasons, they will be doing different things before and after this moment.

I was once asked by prominent neuropsychologist Jeffrey Gray, "What is the difference between two awake individuals, one of them stone deaf, who are both sitting immobile in a room in which a recording is playing a Mozart string quartet?"[24] The answer to that question is, "One of them is hearing the music and one is not." But what does it mean, to a behaviorist, to hear something? To be a normally hearing person means that the

[23] Leonides Castro and Howard Rachlin, "Self-Reward, Self-Monitoring, and Self-Punishment as Feedback in Weight Control," *Behavior Therapy* 11 (1980): 38–48.

[24] As quoted in John Staddon, *The New Behaviorism: Mind, Mechanism, and Society* (Philadelphia, PA: Psychology Press, 2001), 177.

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

present act is part of a series of actions significantly correlated with sounds; to be deaf means that no such correlation exists. The actions (or inaction) of the hearing and deaf individuals at the present moment are overlapping points in two entirely different correlations. Asking Gray's question is like asking, "What is the difference between Picasso and a kindergarten child who both at this moment put a drop of yellow in the upper left corner of a painting?" The difference, of course, lies in what they do before and after this moment.

But behavioral explanations are still unsatisfying. They seem circular—not just teleological but tautological. I claim, though, that behavioral explanations of behavior need not be circular, whereas physiological or cognitive or even spiritual explanations of behavior are often circular. Let me begin to defend this claim with a little bit of history—that of the concept of reinforcement.

The unadorned concept of reinforcement is, in itself, circular: If a reinforcer is made contingent on a particular act, it will increase the frequency of that act. Question: What is a reinforcer? Answer: Anything that increases the frequency of an act it is contingent upon. Attempts to break out of this circularity have defined reinforcement in nonbehavioral terms: a reinforcer is anything that feels satisfying, or anything that reduces a biological need, or anything that reduces a drive, or anything that stimulates a certain area of the brain. However satisfying these definitions may be, none of them has held up (although with the development of magnetic resonance imaging [MRI] machines, the last one is having a comeback). All of these definitions either have fatal exceptions or become circular as satisfactions, needs, or drives rapidly multiply to the point where there is one for each reinforcer. In the end, all of these theories have proved to be unsatisfying—at least for any practical purposes.

The one reinforcement theory that has held up under experiment is David Premack's wholly behavioral theory. According to Premack, all possible actions of a given animal may be ranked in order of value.[25] The ranking between any two actions is determined by a choice test in which the animal has free access to both. The activity the animal spends most time doing is ranked higher in value. Then, if a higher ranked activity is made contingent on a lower ranked activity, the lower activity will increase in rate (be reinforced); if a lower ranked activity is made contingent on a higher one, the higher activity will decrease in rate (be punished). For example, a hungry rat, given a free "choice test" among eating, running in a wheel, and pressing a lever, will choose eating over wheel running and wheel running over lever pressing. Thus, eating is more valuable than wheel running, which is more valuable than lever pressing. With this value hierarchy established by the choice test, the effect of

[25] David Premack, "Reinforcement Theory," in David Levine, ed., *Nebraska Symposium on Motivation* (Lincoln: University of Nebraska Press, 1965), 123–79.

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press

contingencies is predictable: higher valued acts will reinforce lower valued acts; lower valued acts will punish higher valued acts. Consequently, if a lever press serves to release a locked wheel (allowing the rat to run), wheel running will reinforce lever pressing (will increase its rate), while if eating a food pellet serves to motorize the wheel (forcing the rat to run), wheel running will punish eating (will decrease its rate). The degree of reinforcement or punishment will depend on the difference in choice-test value. In other words, behavior in one situation (the free choice test) is used to determine a set of values, which is then used to predict behavior in another situation. Because the choice test, on the one hand, and the contingency, on the other, are independent of each other, Premack's theory is not circular. Value, in Premack's theory, has only a behavioral meaning, not a physiological or introspective one. However unsatisfying it may have seemed initially, this theory has held up under experimental testing and has become quite satisfying to those of us who work with it.[26]

X. Conclusion

From the viewpoint of teleological behaviorism, the autonomy of an act is the degree to which the act is part of a pattern of acts, each done not for its own sake, but for the sake of the pattern as a whole. Autonomy, thus conceived, is not a property of the individual act or a feeling within the actor. It is, rather, a matter of perception—not the actor's perception, but rather the perception made by an observer. But we make such a perception for a reason. Why should we want to discriminate between a person's autonomous acts and those that are compelled? The answer, I believe, is that such a distinction is useful to society in attributing responsibility and in distributing rewards and punishments. I am not saying that there is, or should be, any one-to-one correspondence between autonomy and responsibility on the one hand, or between responsibility and reward and punishment on the other. But I am saying that whether a given act is or is not an autonomous act must be decided on the basis of the needs of an efficient social and political system, and not on the basis of either some property of the particular act or on some state of the actor—spiritual, volitional, cognitive, or physiological—unobservable to other people.

Philosophers are attracted to the concept of internal autonomy because they think that it is a good thing for people to be free to choose their behavior, or free to act without internal compulsion, or free to set moral

[26] Premack's method of determining value is a primitive version of the economist's "revealed-preference" method of determining utility functions; the economist observes choices under two or three sets of constraints, then uses these observations to derive a utility function. The utility function is then used to predict choice under a new set of constraints. If the prediction turns out to be wrong, the economist reexamines the constraints or redetermines the utility function. The method is wholly behavioral. The utility function is just a summary of behavioral observations. It may have no coherent internal representation.

laws for themselves. But why are philosophers attracted to such a concept, and in what sense is autonomous behavior a good thing? Is it good because it just feels good to act freely, or because one internal mechanism rather than another underlies the act? Or is autonomy a good thing because it is useful to society to classify a person's actions as autonomous or not? To argue for the latter is not to imply that autonomy is a wholly arbitrary social classification, having nothing to do with individual behavior. Moreover, autonomous acts are not always good; first-degree murder, after all, implies autonomy. Nor are nonautonomous acts always bad. People may act impulsively (do what they prefer at the moment) and still act well. Nevertheless, as in the case of first-degree murder, it is useful for society to distinguish between discrete acts done for their own sake, or for the sake of immediate rewards, and acts done for the sake of more abstract values and distant rewards. In other words, it is useful for society to distinguish between virtuous or vicious acts that, on the one hand, are performed because they are immediately rewarded (or high in immediate value) and virtuous or vicious acts that, on the other hand, are performed because they are part of a highly valued, temporally extended pattern of acts.

Psychology, State University of New York at Stony Brook

https://doi.org/10.1017/S0265052503202107 Published online by Cambridge University Press