Published in final edited form as:

Drug Alcohol Depend. 2007 September ; 90(Suppl 1): S92–S99.

In What Sense Are Addicts Irrational?

Howard Rachlin

Stony Brook University

Abstract

Rationality is here considered from a functional viewpoint: How may the concept of rationality be best used in talking about addictive behavior? The article considers rationality in terms of overt behavioral patterns rather than as a smoothly operating logic mechanism in the head. The economic notion of rationality as consistency in choice - the property of exponential time discount functions - is examined and rejected. Addicts are not irrational because of the type of time discount function that governs their choices - or even because of the steepness of that function. Instead, rationality is here conceived as a pattern of predicting your own future behavior and acting upon those predictions to maximize reinforcement in the long run. Addicts are irrational to the extent that they fail to make such predictions and to take such actions.

Keywords

addiction; delay discounting; economics; social discounting; rational behavior; utility maximization

1. Introduction

In order to get some sort of handle on what it means to behave rationally, let us look at the concept pragmatically. Instead of asking, “What does rationality really mean?” or, “Where is rationality located?” or, “How does rationality work?” or, “Is this or that animal fundamentally a rational animal?” or, “Is this or that behavior really rational or really not rational?” it might be better to ask, “How should psychologists *use* the word, ”rational?“ What is its proper function in our scientific language?” It seems to me that there are at least two fundamentally different ways to use the term.

Figure 1 illustrates a distinction between two ways of looking at behavior and, by implication, two ways of using the expression, “rational behavior.” The thick vertical line, divides the inside of a person, on the right, from the external world, on the left. On the right is my own version of a generic cognitive model of decision making. I mean it to be an illustration of a *kind* of model and not any particular model. Looked at from the right, information from the outside world comes into the person through the sense organs and enters the cognitive mechanism; the information proceeds through a series of sub-mechanisms; it is perceived, represented (or encoded), processed, used in making a decision, and eventually an overt choice. All of these sub-mechanisms are affected directly or indirectly by memory, by feedback from the external world, and from below by motivational variables such as hunger, thirst or other fundamental or not-so-fundamental drives. Each of the sub-mechanisms may in turn be divided into sub-

Correspondence should be addressed to: Howard Rachlin, Psychology Department, Stony Brook University Stony Brook, NY 11794-2500, howard.rachlin@sunysb.edu, phone: (212) 996-3478

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

sub-mechanisms, and so forth. Cognitive theorists may infer the states of the mechanisms from either non-verbal choice behavior or from verbal reports of those states.

There is some dispute within cognitive psychology whether the lines of division in any particular cognitive theory correspond to, or may be reduced to, physiological mechanisms - that is, whether cognitive psychology is reducible to physiological psychology - or whether cognitive psychology and physiological psychology each carve out non-overlapping units within the nervous system. In any case, the *language* of cognitive psychology does overlap with the *language* of physiological psychology. Our mental vocabulary, including terms such as “memory,” “perception,” and so forth, is common to them both. I therefore call the model on the right side of the diagram a **cognitive-physiological model**. Cognitive and physiological psychologists both have the same ultimate goal - to discover actual mechanisms within the organism (Gazzaniga, 1998).

On the left side of the diagram is what I call the **behavioral-economic model**. This model uses the very same inputs and outputs as the cognitive-physiological model but the boxes stand not for *spatially defined mechanisms* but for *temporally defined contingencies*. A classical contingency is a relationship between two environmental events - the bell and the food powder (in the case of Pavlov’s dogs), or the train whistle and the train. An instrumental contingency is a relationship between behavior and consequences. A fixed-ratio schedule, which says that a rat will obtain a food pellet after every 10 lever presses, the price of a loaf of bread which says that if you give the baker so much money he will give you so much food, the conflicting relationships between smoking and feeling good and between smoking and lung cancer, are all instrumental contingencies.

To contrast the cognitive-physiological viewpoint with the behavioral-economic viewpoint, consider the question, “Why am I sitting here at my computer and typing?” An e-mail came to me several months ago, stimulated my eyes, was perceived, represented, processed, and activated a decision mechanism. The output of that mechanism was a wholly internal decision to write the article. Then that decision activated certain motor centers and actually got me moving. I wrote the due date on my calendar but I also encoded it, or a fuzzy version of it, in my memory. I responded to the various e-mails. Then, when my calendar indicated that the due date was rapidly approaching, my lower motivational processes - what Loewenstein (1996) calls “visceral” processes - became activated - in other words, I panicked - and I began to write. Every word I am writing may be seen in terms of the operation of the cognitive-physiological mechanism inside me. There is no question that some version of a cognitive-physiological theory has to be true. No organism, not even an amoeba - much less a person - is empty.

Yet, there is another way to look at the relation among the heavy arrows of the diagram - from the outside rather than from the inside. For instance, there is another way to look at the question of why I am writing this article. You could say that I am writing because I hope to influence the behavior of readers or at least convince them to look more kindly than they already do on behavioral research. Or, I am writing it because the act of writing down my thoughts will help me to develop my ideas, or because I believe that writing this article will somehow further my career.

The boxes on the left side of the diagram look like the boxes on the right, but they stand for radically different entities. The boxes on the right stand for current states of currently existing mechanisms. You could, in theory, point to them in the same way that you could point to a car’s carburetor. To the extent that the boxes on the right are hidden, they are hidden in space, somewhere inside the organism. On the other hand, the boxes on the left stand for temporally extended contingencies, that is, relationships over time between patterns of behavior and environmental events. You could not point to such contingencies any more than you could point to the relationship between the force of your foot on a car's accelerator and the speed of the car. I call this viewpoint (from the left in Figure 1) "teleological behaviorism."

2. Rational Behavior and Addiction from a Cognitive-Physiological Viewpoint

How, then, should the concept of rationality or rational behavior be used in cognitive theories? For a cognitive theory, rational behavior is usually seen as the product of a smoothly functioning, unimpeded, logical decision mechanism. Different cognitive-physiological theories would have different ideas of how that machine works (i.e., differing *normative* models), and of how it may be impeded. If, as is often the case, the predictions of a given cognitive-physiological theory are disconfirmed, the theory may be modified so as to explain the discrepancy. For instance, a non-linear relationship between probability and decision weight may be assumed, as in Kahneman and Tversky's (1979) prospect theory, or hyperbolic time discount functions may be used instead of exponential time discount functions to predict choice. These changes may be viewed as changes in perception or changes in processing rather than changes in logic, which remains the same as it was in the original theory - the unimpeded operation of an internal logical mechanism. Now, after it has been explained, behavior that was labeled as irrational in the original theory should be labeled rational. This doesn't mean that people have suddenly become smarter than they were before, just that the label applied to their behavior has changed. You might say that the behavior is not so much *rational behavior* as *rationalized behavior*.

Nowhere in this cognitive-physiological conception of rationality (as interpreted by an admitted non-cognitivist, non-physiologist) is the requirement that rational behavior also be conscious behavior. A person might consciously deliberate before acting and still act irrationally or act first and then rationalize her own behavior afterwards or have no conscious awareness of her behavior at all and yet act rationally. Consciousness, whatever it might be (and it might or might not be a useful concept at all within any particular cognitive theory) seems orthogonal to rationality.

Given this very general cognitive view of rationality, how would it apply to addiction? One fairly obvious way is to say that, whereas the addict may possess a perfectly adequate logic mechanism, the operation of that mechanism may be impeded by immediate motivational forces acting from below ("visceral factors"). When such forces are strong enough they may overwhelm the output of the logic mechanism and produce irrational behavior; addictive behavior may then be seen as one form of irrational behavior. If, at a moment when the motivational forces are inactive, you ask a nicotine addict, for example, whether he wants to be a nicotine addict, he says no; this is his logic mechanism working in an unimpeded way. However, when the motivational forces are strong enough - when he is offered a cigarette - he smokes it. An addict may thus differ from a non-addict in two ways: 1. The non-addict's motive to consume the addictive substance may be weaker than that of the addict ("natural virtue" or "temperance," according to Aristotle; for example, my wife hates chocolate) or, 2. The output of the non-addict's logic mechanism may be stronger or less resistant to disruption than that of the addict ("continence," according to Aristotle - what we would call self-control). So, when my wife refuses the chocolate dessert she's being naturally virtuous but not self-controlled; when I refuse it, on the other hand, I'm controlling myself. (Of course, a person may have both natural virtue and self-control.) The concept of "rational addiction" would be self-contradictory from the cognitive-physiological viewpoint since addiction, from that viewpoint, is irrational by definition.[1]

[1] Chorvart and McCabe (2005) recently published an excellent review of the concept of rationality in neurobiology and cognitive decision theory

3. Rational Behavior and Addiction from a Behavioral-Economic Viewpoint

Molar behavioral theories and economic theories of individual behavior (microeconomic theories) take the same form. The instrumental contingencies of behavioral theories correspond to the constraints (prices and budgets) of economic theories. Both specify relationships between behavior and consequences. The behavioral concepts of reinforcement maximization and matching correspond to utility functions of economic theory (Rachlin, 1992).

How does the concept of rationality fit into microeconomics? Economic utility theory assumes that people behave so as to maximize utility under any given set of constraints (prices and budgets in economic language - contingencies or reinforcement schedules in behavioristic language). The object of economic theory is to discover the utility function that actually is maximized. Once such a function is discovered under one set of constraints it may then be tested under another set of constraints. If the same utility function is maximized under both sets of constraints it may then be tested under a third set of constraints, and so forth - until it fails a test as it inevitably must. At that point the utility function is modified or parameters are added to it until it describes behavior under all tested constraints, and the process continues - at least in theory. The end result, in theory, is a grand utility function that is maximized under all possible sets of constraints. This is the method of revealed preference (Samuelson, 1973). Obviously, the desired end point will never be reached - just as no perfect cognitive-physiological model will ever be developed. But in the process of development, economic theory is supposed to become better and better able to predict behavior under any imposed set of constraints. The utility functions are methods the theorist uses to predict a person’s behavior in one situation from his behavior in other situations. Of course there are internal mechanisms underlying all behavior but there may be no particular internal mechanism corresponding to any particular utility function. To use an example from Dennett (1978), a chess player may reliably bring out her queen too soon - that is, bringing out her queen too soon may be an accurate description of her past behavior and useful in predicting her future behavior - but this tendency may not be encoded as such in any specific mechanism inside of her head.

It is often the case that, on the basis of a given utility function, behavior that maximizes utility in the *relatively* short run does not maximize utility, as measured by that particular function, in the *relatively* long run. If there are two people, one of whom maximizes utility in the long run (hence not in the short run) and one of whom maximizes utility in the short run (hence not in the long run), the behavioral-economist could say that the first person’s behavior is (relatively) rational and the second person’s behavior is relatively irrational. An addict would be a good example of a person of this second kind. By smoking, the addict maximizes utility in the short run - the person facing a firing squad might as well have a smoke - but not in the long run - in terms of health, social acceptance, expense, etc.

However, suppose an economic theory claims to have discovered a single utility function that describes both the addict’s and the non-addict’s behavior (Becker and Murphy, 1990). That is, given a particular set of parameters, the addict may be seen to be maximizing utility in the long run. The economic theory would then have rationalized the addict’s behavior. As in the cognitive-physiological theory, this does not mean that people have suddenly become smarter than they were before. It just means that the theory has developed so that, by means of varying the parameters of a single utility function, it can explain and predict both the addict’s and the non-addict’s behavior. A certain consistency is discovered in the addict’s behavior that seemed not to have been there before.

Moreover, as from the cognitive-physiological viewpoint, rationality, from the behavioral-economic viewpoint, has nothing to do with consciousness. Rational addiction means that the addict’s behavior, as well as the non-addict’s behavior, maximizes utility in the long run according to some particular utility function. It does not mean that the future addict sits down on the day of his bar mitzvah and plans out the rest of his life.

4. Crossing Discount Functions and Rationality

Consider the practice of compounding interest by banks. With compounding, simple interest is calculated over some fixed period, *t*, added to the principal, and repeated at intervals of *t*. If, when you came to withdraw your money, the bank calculated simple interest from the time of deposit, you would have an incentive, after a short period, to withdraw your money plus the interest and deposit it in another bank thus compounding the interest yourself. So as not to lose your account in this way, the bank compounds your money for you. As the period of compounding approaches zero, your money would be compounded at an infinite rate, and the resulting overall discount function would approach the exponential function:

$$\frac{v}{V} = e^{-it} \tag{1}$$

where *V* is the balance for an original deposit of *v* after a time, *t*, and with an interest rate, *i*. Now suppose that a delayed reward were discounted using this same function (exponential delay discounting). Then, *v* would be the discounted value of *V*, *v/V* the degree of discounting measured as a fraction, *t* the delay, and interest rate (*i*) a constant representing degree of discounting; with higher values of *i*, discounting would be steeper. That is, the higher the value of *i*, the less a given delayed reward is worth. If the choices of two people in a given situation were described by a delay discount equation, such as Equation 1, and the function describing the choices of one of them had a higher *i*-value than that of the other, that person would be expected to be more impulsive in his choices - would tend to choose smaller-sooner rewards over larger-later rewards to a greater extent - than the person with a lower *i*-value. A theory of addiction based on exponential delay discount functions would say that addicts have higher *i*-values than non-addicts. However, addicts would be no less rational (according to the economist’s definition) than non-addicts - as long as their choices were consistent.

In what sense does exponential time discounting (Equation 1) imply consistent choice? With exponential delay discounting, two delay discount functions with the same interest rate (*i* of Equation 1) would not cross (Ainslie, 1992). If, contrary to fact, exponential delay discounting described actual human and nonhuman choice among delayed rewards, and if a person preferred $100 delayed by 10 days to $95 delayed by 9 days, then, after 9 days had passed, that person would still prefer the $100, now delayed by a day, to the $95 available immediately. In other words, exponential delay discount functions for the $100 and the $95 would not cross. Nevertheless, in many instances, people do change their preferences over time, preferring the larger-later reward when both rewards are relatively distant but switching their preference to the smaller-sooner reward when it becomes imminent. That is, people’s actual discount functions may indeed cross. Does that mean that their choices are irrational?

Because exponential discount functions predict consistency of preference over time, economists call exponential discounting “rational.” In economics, the reason why compound interest functions, such as Equation 1, do not cross is that they are approximated by frequent recalculation of simple interest (compounding). In the case of a savings account, your current balance would be incremented by a constant fraction at frequent periods between deposit and withdrawal. At any instant, all current balances (with the same interest rate) are increased by the same percentage of their current principal; thus, a lower balance can never catch up to a higher one on the basis of interest alone. (It is as if all employees of a company always received the same percentage raise regardless of their term of service or their performance. Under such conditions the rank order of their salaries would never change.)

While it is common for *banks* to approximate exponential discounting by compounding interest over the period of a loan when borrowing or lending money for an indefinite period, it is not common for an *individual* to compound the appreciation of value continuously over fixed delays when choosing among delayed rewards. Indeed, there is strong evidence that people's choices are not well described by the exponential discount function of Equation 1. The choices of people as well as nonhuman animals have been found to conform instead to *hyperbolic discounting*.

With hyperbolic delay discounting:

$$\frac{v}{V} = \frac{1}{1 + kD} \tag{2}$$

where *v/V* is the degree of discounting measured as a fraction, *D* is delay of reward (corresponding to *t* in Equation 1), and *k* is a constant measuring degree of discounting (corresponding to *i* in Equation 1). In virtually all experimental determinations of psychological discounting, Equation 2 (or a variant with the denominator exponentiated) is the form empirically obtained (Green and Myerson, 2004).

Reconsider the above example in which a person is supposed to prefer $100 delayed by 10 days to $95 delayed by 9 days. With Equation 2 (and a sufficiently high value of *k*), after 9 days had passed, that person would now prefer the $95 available immediately to the $100, now delayed by a day. In other words, unlike exponential delay discount functions with the same interest rate (*i*), hyperbolic discount functions with the same discount rate (*k*) may cross. Despite the reversal in preference as time passes, there is nothing necessarily irrational about hyperbolic discounting per se, even by forward-looking organisms. That is, crossing functions are not in themselves necessarily irrational.

The hyperbolic discount functions of two- and three-dimensional energy propagation may cross under conditions corresponding to crossing delay discount functions (For example, the sound energy from a radio close to your ear may be more intense than that of a moving subway train 10 feet away but, stepping back 10 feet - now you'd be 10 feet from the radio and 20 feet from the train - the sound energy intensities of the two would reverse.) But there is nothing irrational about this purely physical process.

What may be considered irrational, however, is a failure to account for a change of mind when it is known that a second choice will be offered and, on the basis of past experience, that one's own preference will reverse (O'Donoghue and Rabin, 1999). The behavior of an alcoholic who vows, during a morning hangover, never to drink again, and then goes to a party, when in the past he has always gotten drunk at such parties, may be labeled as irrational - not because the addict changed his mind about drinking when he got to the party but because the addict failed to anticipate that he would change his mind.

5. Rationality In Social Behavior

The following passage from Anthony Trollope's novel, *The Way We Live Now* (1875/1982, Oxford: Oxford University Press) presents an analogy between a character's selfishness and his impulsiveness (p. 17):

> Whether Sir Felix ... had become what he was solely by bad training, or whether he had been born bad, who shall say? It is hardly possible that he should not have been better had he been taken away as an infant and subjected to moral training by moral teachers. And yet again it is hardly possible that any training or want of training should have produced a heart so utterly incapable of feeling for others as was his. He could not even feel his own misfortunes unless they touched the outward comforts of the

> moment. It seemed that he lacked sufficient imagination to realize future misery though the futurity to be considered was divided from the present but by a single month, a single week, - but by a single night.

Trollope here attributes Sir Felix's selfishness, his social narrowness, to his lack of self-control, the narrowness of his time horizon. Is there a relationship between rationality in self-control and rationality in social-cooperation?

Recently, Jones and Rachlin (in press) found that Equation 2 precisely described the average results of more than 300 human participants who each chose whether to share a hypothetical amount of money with another person at a greater or lesser social distance; the discounting variable, social distance (*N*), was measured as numerical order in closeness, to the participant, of the person who would be sharing the money (#1 being the closest, #2 being the second closest, and so forth). Participants were asked to imagine that they had made a list of the 100 people closest to them in the world ranging from their dearest friend or relative at #1 to (possibly) a mere acquaintance at #100. Then they were given a series of two-column lists. Column A listed dollar amounts ranging from $75 to $155 in $10 increments "for you alone" (the "selfish" option); for each column-A amount, column B repeated the same alternative: "$75 for you and $75 for the [*Nth*] person on the list" (the "generous" option). Participants chose between A and B for each item on each list. The social distance from the participant to the recipient (*N*) was constant on each list but varied between lists. The dependent variable was the crossover point between the "generous" alternative (preferred with low column-A amounts) and the "selfish" alternative (preferred with high column-A amounts). This crossover point represents the maximum amount of money the participant was willing to forgo in order to give $75 to person-N.

We found that the greater the social distance of the receiver from the sharer (*N*), the less money the sharer was willing to forego. That is, "generosity" was discounted by social distance according to the hyperbolic formula:

$$\frac{v}{V} = \frac{1}{1 + kN} \tag{3}$$

where *v/V* is the fraction of their own reward our participants indicated they were willing to forgo in order to give $75 to person-N. With social distance (*N*) taking the place of delay (*D*) in Equation 2, and $k = .05$, the variance accounted for (R^2) by Equation 3 was .997. Thus, delay and social discounting are found to take the same form.

Equation 3 was fit to the median crossover points of all the subjects. Fitting the equation to individual participants, the constant (*k*) varied over a wide range. Just as a high delay-discounting constant (*k*) in Equation 2 implies lack of self-control, so a high social-discounting constant (*k*) in Equation 3 implies lack of generosity. In as yet unreported data from our laboratory, there was a significant correlation (among college students) between reported number of cigarettes smoked per day by an individual and that individual's social-discounting constant. Moreover, individual delay discounting was significantly correlated with social discounting over individual participants. That is, people whose discount functions implied a tendency to be self-controlled also tended to be generous to others. Why might this be so?

As Ainslie (1992) and Rachlin (2000) point out, you can view an individual over time analogously to a series of individuals over space. Again let us look to the arts for an example - the *Seinfeld* TV show this time. In one of his introductory routines (as I remember it) Seinfeld talks about "night-Jerry" and "day-Jerry." Night Jerry has fun; he stays out late, gets drunk, spends money. Day-Jerry is the one who has to wake up early in the morning with a painful hangover. Night-Jerry has only contempt mixed with pity for day-Jerry while of course day-Jerry hates and resents night-Jerry. But what can day-Jerry do to get even? His only recourse, Seinfeld says, is to stay in bed late, perform badly at his job, and then get fired; then night-Jerry won't have enough money to go out and get drunk. Jerry's problem is very much like Sir Felix's problem in Trollope's book. Both characters fail to identify with their future selves. Their delay-discount functions are very steep because they see their own selves at a future time as socially distinct from their own selves at the present moment. Addicts, I would argue, are in such a position.

Non-addicts, on the other hand, distinguish very easily and quickly between their own future selves and other people. Brown and Rachlin (1999) attempted to compare self-control and social cooperation with humans playing a game. The game could be played either by a single player ("alone") to study self-control, or by a pair of players ("together") to study social cooperation. The participants were all female Stony Brook undergraduates.

The game board is diagramed in Figure 2. It consisted of a rectangular plastic tray divided into 4 compartments ("boxes"). Each box contained 3 items:

- A red or green index card with a picture of a door ("red doors" or "green doors")
- A red or green key
- 1, 2, 3, or 4 nickels

The upper boxes both contained red doors; the lower boxes both contained green doors. The left boxes both contained red keys; the right boxes both contained green keys. Note that each right box held 1 more nickel than the box to its left, and each upper box held 2 more nickels than the one below it. All the items in the boxes were visible to the players.

5.1. The self-control game ("Alone")

Each trial began with the apparatus as pictured in Figure 2. To start, a player was given a red key. The player could use that key to "open" one or the other red door (to choose either the upper left or upper right box). The used key was then surrendered. If the upper left box was chosen, the player was permitted to take the 3 nickels and the red key from that box. If the upper right box was chosen the player was permitted to take the 4 nickels and the green key from that box. Then the nickel(s) and key taken were replaced by the experimenter and the next trial began. If a red key had been received on the previous trial, the player could again choose between the two red doors as before; if a green key had been received on the previous trial, the player could use the key to "open" one or the other green doors (to choose a key and nickels from either the lower left or lower right box).

The alone game is a self-control procedure in the sense that the behavior leading to the higher current reward (choosing the right box with 2 or 4 nickels plus a green key) conflicted with the behavior that maximized overall reward (choosing the left box with 1 or 3 nickels plus a red key). Choosing the right box always earned the player one more nickel than choosing the left box did, but at the cost of obtaining a green key. With the green key the player paid for the 1-nickel gain (for choosing the right box) on the previous trial with an average 2- nickel loss (having to choose between the lower boxes) on the present trial.[2]

The best overall strategy in the alone game is to always choose the left box, always receive a red key, and always earn 3 nickels. Always choosing the right box yields an average return of

[2] The alone version of the game duplicates the contingencies of a prisoner's dilemma game against an opponent playing "tit-for-tat." Tit-for-tat says, cooperate on this trial if your opponent cooperated on the previous trial and defect on this trial if your opponent defected on the previous trial. Thus, a player playing against tit-for-tat who cooperates on the present trial will be able to choose next time between the higher rewards (because the other player cooperated). The player against tit-for-tat who defects on the present trial will be forced to choose next time between the lower two rewards (because the other played defected). These are the very contingencies set up by the keys and doors of the alone condition.

2 nickels. Alternating between the left and right boxes yields an average return of 2.5 nickels [(3 + 2)/2 or (4 + 1)/2]. Only on the very last trial does it pay to choose the right box but the subjects did not know when the experiment would end.

Because current choice of the lower available reward always leads to a higher next-trial reward, current choice in the alone game depends on the degree to which the (higher) next-trial reward is discounted. Because it cannot be obtained until the next trial, the higher future reward may be discounted by delay. But another possible source of discounting is social discounting. A player may currently discount higher future reward by the social distance between herself now and herself in the future.

5.2. The social cooperation task (“Together”)

The game, as played by two players together was the same as when played alone, except the two players, playing on a single game board, made choices on alternate trials. After using her key to open a box, each player took the nickels in the box for herself but then handed the key to the other player. Thus, after the first trial, whether a player was permitted to choose between the upper boxes (3 or 4 nickels) or between the lower boxes (1 or 2 nickels) depended on the other player’s choice on the previous trial. This task has the essential properties of a prisoner’s dilemma game.

Playing this game together, income would be maximized (at 3 nickels per trial) for each player if both players repeatedly chose the left box (cooperated). However, the individual player would always gain more on the present trial by choosing the right box (defecting). The penalty for defecting, of having to choose between the lower boxes, is suffered not by the player who defects but by the other player, who inherits the green key.

Cooperating is the very worst strategy in the together game, unless the other player also cooperates. Therefore the only reason to cooperate (within the demands of the game) is to influence the other player to cooperate subsequently. The reward for cooperating in the social cooperation version of the game must be discounted not only by the delay to the player’s next turn but also by the probability that the other player will reciprocate. Most people’s estimation of the probability of other people’s future cooperation might be expected to be lower than their estimation of the probability of their own future cooperation. For this reason, a player who cooperates with her own future self in the alone game (who consistently chooses the lower current-trial reward) may defect from the interests of her partner in the together game. This was exactly what Brown and I found. Participants who learned to cooperate at a relatively high rate in the alone game suddenly defected as soon as they began playing in the together game. The drop from cooperation (with themselves) to defection (with another player) was both sharp and sudden. *Experience with self-control seemed to have no effect on social cooperation*. The participants in this experiment easily learned to cooperate with their own future selves, albeit over a short time span, but that learning did not transfer to cooperation with another person over that same time span.

Addicts fail to cooperate with their past selves; this failure lowers their estimation of the probability that their future selves will cooperate with them. Addicts treat their future selves as they would treat people far from them in social distance - people unlikely to reciprocate any cooperative behavior they may exhibit. If, in future interactions, other people are unlikely to cooperate with you, regardless of your behavior now, it would be foolish to cooperate with them now. To take an example from Aristotle, an addict is in the position of a soldier in a rout in battle who, if he turns and makes a stand, will be ignored by the other fleeing soldiers. The irrationality of addicts may be most usefully understood to lie, not in the description of their choices by one sort of discount function or another, but in their failure to identify with their past and future selves - their failure, that is, to distinguish between their own past and future selves, separated from them in time, and other people, separated from them in social space. That is, even when she is nominally playing an alone game - a game against her future self - the addict plays it as a non-addict would play a together game - as a game against other people, people socially distant from her.

6. Implications For Treatment of Addiction

Although the purpose of this article is to outline and discuss two viewpoints (cognitive-physiological and behavioral-economic) of the concept of rationality in addiction, and not to suggest treatment methods, it may be clarifying to consider how the two viewpoints result in two different approaches to treatment. The cognitive-physiological viewpoint would focus on the two internal sub-mechanisms of addiction - the logic mechanism and the motivational interference with that mechanism. Current neurobiology has given us hints of where in the brain these internal mechanisms and interactions are located but we know very little about how they actually work. Moreover, in its focus on internal mechanisms, the cognitive - physiological approach tends to ignore behavioral context. Cognitive-physiological treatment would focus on drugs and perhaps other physiological manipulations. Behavioral treatments, from the cognitive-physiological perspective, would be designed not primarily to alter behavior itself but to alter the state of internal mechanisms. In any case, the focus is always on the addictive behavior itself rather than on contextual factors.

The behavioral-economic viewpoint, on the other hand, gives equal weight to the target behavior and its context. It suggests not only increasing the price of the target behavior but also reducing the price of economic substitutes. In the case of addictive behavior, social interaction and contingent social approval have been found to be economic substitutes for smoking, cocaine addiction, heroin addiction, and other addictions (Rachlin, 2000, Ch. 4). Under certain conditions, manipulation of the price of substitutes is more effective in controlling addiction than manipulation of the price of the addictive substances themselves.

Moreover, the behavioral-economic viewpoint suggests training addicts in self-control by increasing the temporal extent of patterns of alternative behavior. Learning an effective personal rule that may come to control and reduce addictive behavior is not primarily a cognitive problem. It is easy enough for a person to learn to repeat a rule. The difficult task is to bring behavior under control of that rule (as a discriminative stimulus signaling valuable non-addictive patterns). For this task, behavioral methods are currently superior to cognitive or physiological ones. Nevertheless, pursuit of both cognitive-physiological and behavioral-economic treatment approaches would seem to be the most fruitful tactic for treatment.

7. Rationality and Self-Control in Everyday Life

You are on the road in a place far from home - a place where, the odds are, you’ll never return again. You stop at a restaurant to eat. Should you leave a tip? Given that the point of leaving tips at a restaurant is to reinforce the efficiency and courtesy of the server so that when you come there again you will receive good service, or at least not deliberately bad service, a logic machine would say *no*, or at least that you should leave the minimum amount that would avoid a scene. Yet most people do leave tips in these situations, and more than a minimal amount. Is such behavior irrational?

Looked at from my use of the phrase, “rational behavior” to mean behavior by an individual in the present that takes that same individual’s future behavior into account, leaving tips in restaurants where you will almost certainly never return may indeed be rational; moreover, not leaving a tip may be irrational. How can this be so?

Consider the person who, attempting to be rational in the cognitive-physiological sense, decides each time he pays a restaurant bill whether to leave a tip. Not only will that person have to pay attention to the quality of the service and the probability of his coming back to that particular restaurant and to process that information, but he will have to consider the probability that, faced with adding a not inconsiderable sum of money to his bill, "visceral factors" might influence his reasoning and bias him in the direction of leaving less than the optimal amount. I am not aware of research on the subject but my guess is that, when evaluating the amount of money to be spent at some future time in some specific future situation, most people will calculate an optimal amount greater than the amount they calculate should be spent in that very same situation right now. In other words, "visceral factors" (acting currently but not on contemplation of the future) influence not only our motives but also our reasoning.

Realizing that our on-the-spot decisions are often biased in this way, we allow our behavior to conform to some molar pattern that we have found to be generally optimal - always leave a tip (or never leave a tip, or leave 15 ± 2% depending on how we feel, if that's what we have determined). Such behavior, accounting as it does for our own frailty, may usefully be labeled as more rational than that based on a more precise calculation of current contingencies that fails to account for that frailty. Similar considerations may govern other social activities such as contributing to public broadcasting or to charities, voting, not littering, bussing your tray in a cafeteria, etc.

Just as it may be rational in social-cooperation situations to obey general rules - to make decisions on a global rather than on a local basis -- it is rational to obey general rules in self-control situations. The addict may label his own behavior as rational when he lights up one "last" cigarette or drinks one "last" drink; after all, there will be great pleasure and virtually no harm in it if this one is indeed the last. But, as we all know, this is just the opposite of rational behavior - especially for an addict. To achieve self-control, addicts and non-addicts alike must avoid making each decision on an apparently rational case-by-case basis and learn to make particular decisions in conformity with optimal molar patterns (Rachlin, 2000).

Acknowledgements

The writing of this article and the experiments described therein were supported by a grant from The National Institute of Mental Health.

References

Ainslie, G. Picoeconomics: The Strategic Interaction of Successive Motivational States Within The Person. Cambridge University Press; New York: 1992.

Brown J, Rachlin H. Self-control and social cooperation. Behavioural Processes 1999;47:65–72.

Chorvart T, McCabe K. Neuroeconomics and rationality. Chicago-Kent Law Review 2005;80:1235–1254.

Dennett, D. Brainstorms: Philosophical Essays on Mind And Psychology. Bradford Books; Montgomery, VT: 1978.

Gazzaniga, MS. The Mind's Past. University of California Press; Berkeley: 1998.

Green L, Myerson J. A discounting framework for choice with delayed and probabilistic rewards. Psychological Bulletin 2004;130:769–772. [PubMed: 15367080]

Jones B, Rachlin H. Social discounting. Psychonomic Bulletin and Review 2006;17:283–286.

Kahneman D, Tversky A. Prospect theory: An analysis of decisions under risk. Econometrica 1979;47:263–291.

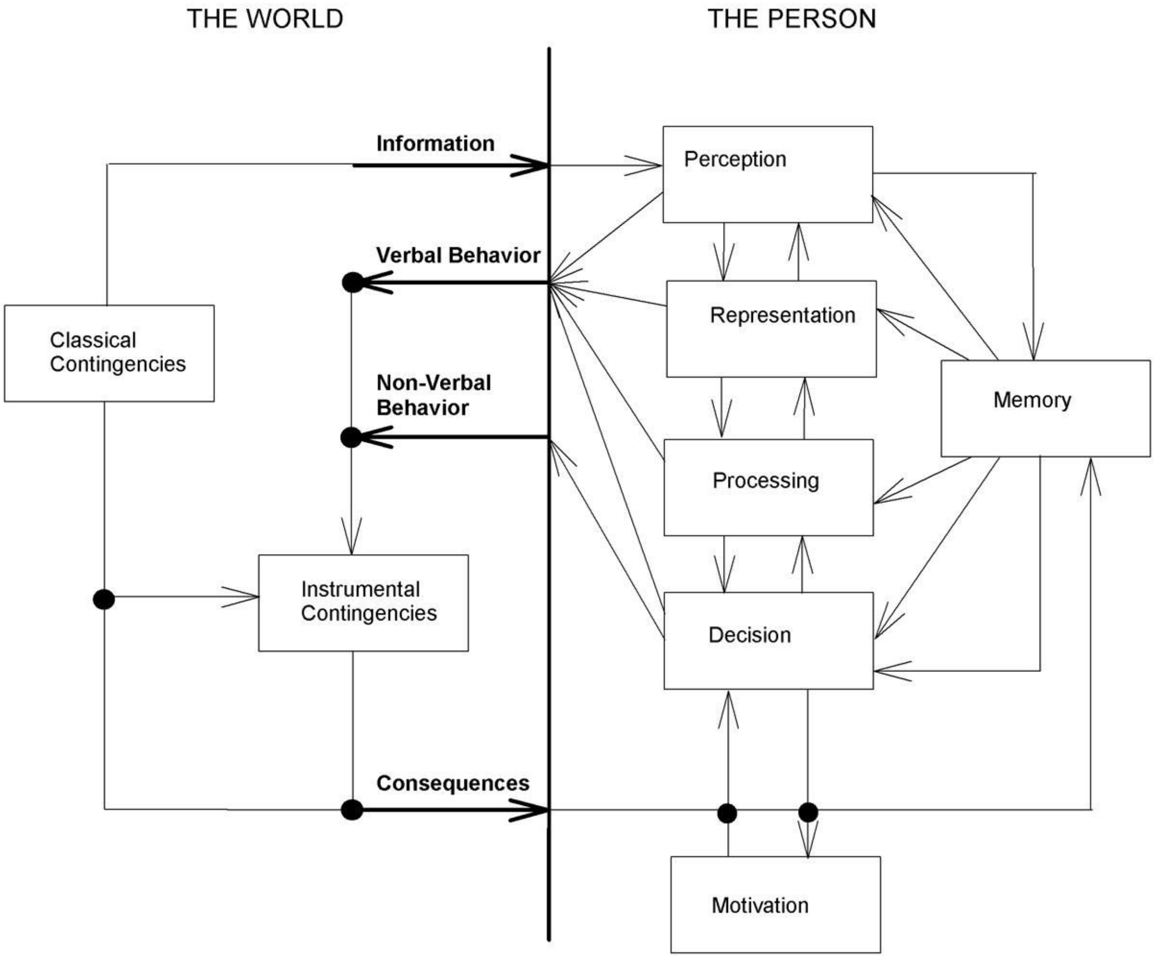
Loewenstein G. Out of control: Visceral influences on behavior. Organizational Behavior and Human Decision Processes 1996;65:272–292.

O'Donoghue T, Rabin M. Doing it now or later. American Economic Review 1999;89:103–124.

Rachlin H. Teleological behaviorism. American Psychologist 1992;47:1371–1382. [PubMed: 1482004]

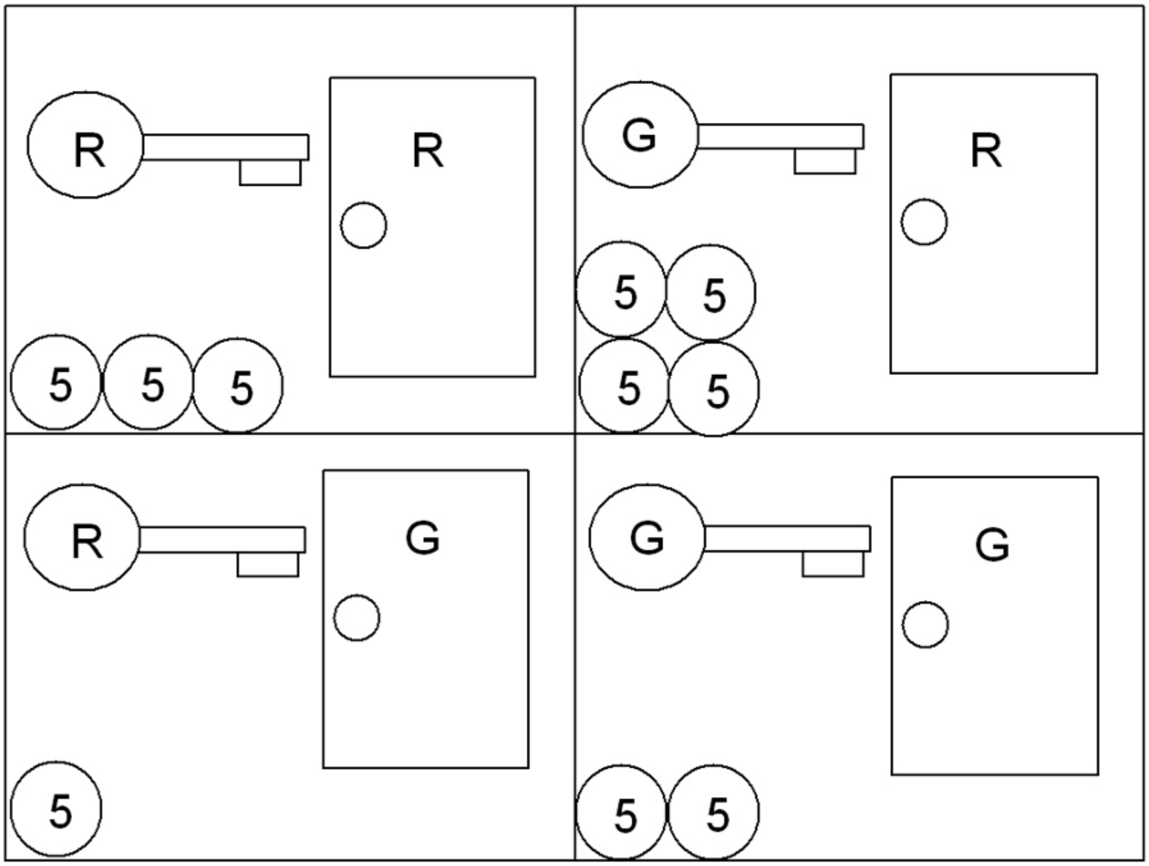
Rachlin, H. The Science of Self-Control. Harvard University Press; Cambridge, MA: 2000.

Samuelson, P. Economics: An Introductory Analysis. 9th ed.. McGraw-Hill; New York: 1973.



1.

A person’s mind as seen from a cognitive-physiological viewpoint (to the right of the vertical line) and from a teleological behavioral viewpoint (to the left of the vertical line). From the cognitive-physiological viewpoint mental terms refer to mechanisms inside the person. From the teleological behavioral viewpoint mental terms refer to patterns of interaction between the whole person and the environment over extended time periods.



2.

Diagram of the game used in Brown and Rachlin's (1999) experiment. A plastic tray was divided into four compartments each containing a number of nickels, a red or green key and a picture of a red or green door. Players who had a red key could "open" a red-door compartment and obtain the nickels in that compartment plus a key for the next round; similarly, players who had a green key could "open" a green-door compartment and obtain the nickels and key in that compartment. The game duplicates the contingencies of a repeated prisoner's dilemma in that repeated choice of the compartment with the lesser number of nickels obtains the highest reward in the long run when playing the game "alone" or the highest group reward when playing "together."