JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 2010, **94,** 95–111 NUMBER 1 (JULY)

# *HOW* SHOULD *WE BEHAVE?* *A REVIEW OF* REASONS AND PERSONS *BY DEREK PARFIT*

HOWARD RACHLIN

STONY BROOK UNIVERSITY

In this book Parfit attempts to develop a rational, non-religious ethics. Instead of asking, ''What does God tell us to do?'' he asks, ''What does reason tell us to do?'' Given a set of simple assumptions, he considers whether it is possible to be consistently selfish or consistently good. Analyses of personal dilemmas (problems of self-control) and moral dilemmas (problems of social cooperation) show that neither consistent selfishness nor consistent goodness is logically possible. Instead, a fine balance must be maintained between, on the one hand, our immediate versus long-term good and, on the other, our own good versus that of other people. Ultimately Parfit fails to develop a formula by which such a balance may be struck. Parfit's analysis is consistent with behavioral analysis in its reductionistic view of the self and the parallel it draws between relations with other people and relations with oneself at other times. Parfit's analysis is inconsistent with behaviorism in its view of the mind as internal, available to introspection, and able to cause behavior. His nonfunctional mentalism leads Parfit to inconsistencies and blocks the path to a consistent ethics. Teleological behaviorism's view of the mind in terms of patterns of overt behavior is not hampered by these inconsistencies and may lead to a functional rather than purely rational ethics.

*Key words:* behaviorism, ethics, mind, morality, prisoner's dilemma, rationality, self, self-control, teleological behaviorism

---

Although Parfit's *Reasons and Persons* (1984) was published a quarter of a century ago, I believe that most behavior analysts are unaware of its existence, let alone its importance for behavior analysis.[1] Its goal (explicitly stated only at the very end) is to develop a ''non-religious ethics.'' In the last paragraph of the text (p. 454) Parfit says:

> Disbelief in God, openly admitted by a majority, is a recent event, not yet completed. Because this event is so recent [not even within view in the U.S.], Non-Religious Ethics is at a very early stage. We cannot yet predict whether, as in Mathematics, we will all reach agreement. Since we cannot know how Ethics will develop, it is not irrational to have high hopes.

From this you can see that Parfit does not claim, after some 550 tightly-packed pages including 10 appendices and 28 pages of notes, to have reached his goal.[2] Nevertheless, his attempt to take an ''impersonal'' approach to ethics has aspects in common with behaviorism. Behaviorism is often accused, falsely and perversely, of being an immoral psychology. This book, in attempting to find a rational basis for non-religious morality, provides a basis also for a behavioristic morality.

However, Parfit's ultimate criteria are introspective. His method is to start with a set of abstract moral principles and then to compare them with each other by drawing out their logical implications. So far so good. But, at each stage, the logical structure is tested not against its function in human life but against its intuitive plausibility. Parfit repeatedly builds elaborate theoretical structures out of his axiomatic principles and then stands aside,

---

This review was prepared with the assistance of NIDA Grant DA02652021.

Address correspondence to Howard Rachlin, Psychology Department, Stony Brook University, Stony Brook, NY 11974-2500 (e-mail: howard.rachlin@sunysb.edu).

doi: 10.1901/jeab.2010.94-95

[1] I myself became aware of it only a few years ago. An anonymous reviewer of Rachlin (2002) wrote that Parfit contradicted my argument that self-control is closely tied to social cooperation. When I read the article by Parfit to which the reviewer referred I discovered the exact opposite to be the case.

[2] Amazingly, the book has no index of terms and no glossary. The following passage is not atypical (p. 401): ''The Wide Average Principle restates in person-affecting terms the Impersonal Average principle. We need a new principle of beneficence. We need a principle that both solves the Non-Identity Problem and avoids the Repugnant Conclusion. We may also want this principle to explain the Asymmetry.'' Once it is unpacked, this sentence makes sense. To find definitions of these terms, however, the reader must scurry back through the book – unless, as I eventually did, she writes down each new term and its definition on an index card as she comes across it.

wringing his hands as it were, while the structure crashes to earth on the grounds of intuitive implausibility.

For Parfit, the mind consists of a series of ''psychological states'' or ''mental events.'' We are happy for a while, sad for a while, experience pleasure or pain for a while, have aims and goals that may be highly abstract, have motives and desires, and make plans—and these states persist or disappear. Our minds are nothing but their succession. Good and bad, for Parfit, are measured in terms of these states—desires, pleasures, pains, hopes, fears, etc.—occurring presumably inside our bodies, overlapping with each other, coming and going.

Mental events, for Parfit, may be caused by (but are not identical to) brain events; they may interact with each other; we can report when we have them (i.e., introspect); they may cause us to act. The bottom-line issue for Parfit's morality is the quality and quantity of such states or events—our own, those of other people born and unborn, those existing at the present time, those that existed in the past, and those potentially existing in the future. Behavior counts for Parfit only to the extent that it may affect the quality and quantity of these states or events.

Pain and suffering count for Parfit only in their immediate effects. A distant pain is bad only to the extent that it will actually be experienced. He says (p. 180), ''Pains matter only because of what they are like when they are in the present....This is why we must care more about our pains when we are *now* in pain.'' Parfit ignores pain's *function*—to escape and avoid bodily damage. Pain is bad, I believe, only when its natural function is bypassed or perverted (Rachlin, 1985). The *reason why* we have pain is because pain is closely tied to bodily damage, and we have evolved to escape or avoid bodily damage—to ourselves (as in pulling your hand from a hot stove) and to others (if a baseball player falls and writhes on the field, the game stops and people rush to his aid). Parfit ignores such issues.

The same sort of problem arises about mental states other than pain, such as beliefs. Parfit says (p. 120), ''Besides reasons for believing, there are reasons for acting.'' For him, a belief (together with another mental state, a motive) may cause an action, but is, in its nature, ''beside'' actions, something entirely different from actions. It is a disposition, where a disposition is, for Parfit, like a pain—a mental state conceived as continuously present while actions are going on and between actions. To take Parfit's example (p. 120), consider claustrophobia. A woman has in the past avoided and escaped from closed-in spaces more frequently and with more violence than do most people. She acts panicky in those situations and says that they terrify her. All observers, including she herself, agree that she is claustrophobic. The behaviorist's view of such behavior is what Parfit would call ''reductionist.'' A behaviorist would say that her claustrophobia is *nothing but* such a pattern of verbal and nonverbal behavior. The woman's particular claustrophobic acts are, to her claustrophobia, a behaviorist would say, as the notes of a melody are to the melody itself. Although Parfit's view is (admirably) reductionist regarding the concept of self (I will discuss this presently) his view is (unfortunately) nonreductionist regarding claustrophobia. Claustrophobia for him is a psychological state, a *disposition* conceived by Parfit as something other than her long-term pattern of behavior. What that other something is, is hard to say; it could not be neural activity alone. Like behaviorists, Parfit refuses to identify mental acts with neural activity. How interesting it would have been if Parfit could have applied his reductionism to people's *minds* as well as to their *selves*. That, as we will see, would have brought him closer to behaviorism.[3]

Although Parfit does not explicitly argue against behavioral theory, there is an implicit antibehavioral argument in his moral reasoning. We may, Parfit says (p. 121), ''...have a reason to act morally even if we have no desire to do so.'' Two people both may be acting morally, one with the desire to act morally and one with no desire to act morally but with sufficient reasons to do so. Parfit might say that, if such a scenario is possible, behaviorism must be wrong. He could claim that, because two people may act similarly but for different reasons, reasons (conceived by Parfit as causal mental states) cannot be the same as (or

[3] For a behavioral and functional conception of mental states see Rachlin (1994). For a behavioral and functional conception of pain specifically see Rachlin (1985).

nothing but) actions. This sort of argument merely takes the typical philosophical objection to behaviorism to a different level. To quote myself, here is that argument and its answer (Rachlin & Frankel, 2009, p. 133):

> From the neuroscientist Jeffrey Grey (as quoted by John Staddon, 2001, p. 177): ''What…is the difference between two awake individuals, one of them stone deaf, who are both sitting immobile in a room in which a record player is playing a Mozart string quartet?'' The obvious answer is tautological: one is hearing the quartet and the other is not. But what does it mean to hear something? For Gray and many philosophers, even some modern ones…it means that something nonphysical is going on in their heads—the hearing person is having a ''quale'' or a ''train of sensations'' or a ''raw feel'' and the deaf person is not…. For the psychologist the difference between a hearing person and a deaf person lies in their discriminatory behavior over time. The two people in the room with the quartet playing are…both doing the same thing in the short run but doing different things in the long run. Gray's postulation that one can hear and the other is stone deaf means that for one person sounds are discriminative stimuli while for the other they are not. The identity of their behavior at the present moment means no more than…the identity of the behavior of Picasso and a kindergarten child both, at the present moment, painting a yellow line in the upper right corner of a piece of paper. In all cases the crucial distinction lies in behavior in the long run, behavior considered in molar terms, behavior as real as a punch in the nose….
>
> In order to treat a pattern of behavior over time as a real event, and to deal with it as a scientific datum, it is not necessary to postulate a spiritual event, a neurophysiological event, or a cognitive event whose only property is to internally represent the behavioral pattern at every moment during its performance….It is these internal representations, not the molar patterns they supposedly represent, that are… made up and not real.

The difference between deaf and hearing people, both doing the same thing at the present moment, lies in the behavioral context of the present moment. Similarly, the difference between a person acting morally (over some period of time) for good reasons and one acting morally (over a corresponding period) simply because he desires to do so lies in the still wider or more abstract patterns of which these subpatterns form a part.[4]

One of the great virtues of this book is that it shows how apparently simple principles like: ''(C1) There is one ultimate moral aim: that outcomes be as good as possible'' (p. 24), may be impossible to follow with perfect consistency. Nevertheless, Parfit would not want to say that, because he finds no entirely rational ethics, it is impossible to act morally for reasons (as opposed to desires). Except at the extremes, the distinction between reasons and desires is indeed meaningful. For Parfit, this distinction is between different internal efficient causes of behavior (between one and another ''psychological state'' or ''mental event'' observable only by introspection); for a teleological behaviorist the distinction is between wider, more abstract behavior patterns and narrower, more particular ones (observable, with enough time, by anyone).

Alright then, you may be asking, why should a behaviorist be interested in this book? For two reasons, mainly, in addition to the goal of establishing a secular morality. First, as indicated above, Parfit makes a compelling case for what he calls *reductionism*, which I will call *self-reductionism*: ''Each person's existence just involves the existence of a brain and body, the doing of certain deeds, the thinking of certain thoughts, the occurrence of certain experiences, and so on'' (p. 211). Individual actions and thoughts take time and overlap with each other giving us the illusion of a continuous self. But, aside from that overlap, there is no ''further thing''—either spiritual or physical—to a person's self. You may have as little in common with yourself twenty years from now as you do currently with a distant cousin. A person's continuous character is, according to

[4] The distinction between acts due to desire and acts due to reason is like Aristotle's distinction in *Nicomachian Ethics* between ''natural virtue'' and ''continence'' (Rachlin, 1994, pp. 106–107). My wife, who hates chocolate, would be naturally virtuous when she refuses a chocolate dessert while I, loving chocolate, would be continent even though we are identically refusing dessert at this moment. The difference between her hatred of chocolate and my love for it is not essentially a difference in our internal states (although they are undoubtedly different) but a difference in our long-term patterns of overt behavior. It is no odder that identical acts occur within different behavioral patterns than it is that identical notes occur within different melodies or identical sequences of notes in different musical works.

Parfit, no more (or less) real than the character we ascribe to groups of people such as families, clubs, firms, or nations. As he says (p. 211): ''Most of us are reductionists about nations….Nations exist. Though nations exist, a nation is not an entity that exists separately, apart from its citizens and its territory.''

If your future self is no closer to your present self than is another person, then there should be no essential difference between your relations to your future self and your relations to other people. Since your concern or lack of concern for other people involves *moral* issues, Parfit says, so does your concern or lack of concern for your future self. Therefore (and this is the second great virtue of Parfit's book) issues of social cooperation and altruism on the one hand and self-control on the other are treated in the same way. A motorcyclist's refusal to wear a helmet would be, for Parfit, a moral issue, not just because other people may have to pay a price for her brain injury, but also because her future self may have to pay a price.

As Parfit emphasizes, both self-reductionism and the correspondence of altruism and self-control are difficult concepts for people to accept. It is an irony that while Parfit says that his own belief in self-reductionism has been good for him, and claims (correctly) that people would be much happier if only they could accept self-reductionism, he is blind to similar beneficial possibilities of what I will call, *mind-reductionism*—the notion that there is no ''further fact'' to mental acts or psychological states than temporally extended patterns of overt behavior. Just as a nation is nothing more than its territory and people, and a person is nothing more than his body, his behavior, and his psychological states, so (the teleological behaviorist says) his psychological states (sensing, perceiving, remembering, thinking, feeling, and so forth) are no more than his behavioral patterns.

To define mind-reductionism in Parfit's terms: *Mind-reductionism is the notion that each person's mind just involves the existence of a brain and body, and the doing of certain deeds.* Consider the following scenario (mine, not Parfit's):

> *Abe is on his deathbed. Throughout his married life of 50 years he has alternately beaten and neglected his wife and children. They hate him. But he tells them that whenever he has reflected and introspected, and he has done this frequently over the years, ''deep down'' he has always loved them.*

Can Abe possibly be right? For Parfit, and almost all modern philosophers, it is at least conceivable that he is right. For them, such a disconnect between love and behavior may be an *empirical* rarity but is nevertheless not impossible because, for them, love is an internal event potentially accessible to introspection. For a *mind*-reductionist, on the other hand, it is as inconceivable that Abe loved his family as it is that a pair of teams playing baseball are really playing football. For a mind-reductionist, love, like other mental states, *is* overt behavior. It is thus impossible for love to go one way and a long-term pattern of behavior to go another way.[5]

Like some of the individual and moral theories he rejects, Parfit's self-reductionism is a mixed theory. It is reductionist when it comes to the self, but anti-reductionist when it comes to the mind. Nevertheless it is easy to translate his language into behavioral (i.e., mind-reductionist) language. Here is what he says (p. 251) about thought, with behavioral translation in brackets:

> The existence of a thinker just involves the existence of his brain and body, the doing of his deeds [his particular actions], the thinking of his thoughts [a certain type of behavioral pattern], and the occurrence of certain other physical and mental events [other behavioral patterns]. We could therefore redescribe any person's life in impersonal terms.

If the thinking of a person's thoughts occurs wholly or in part within the person, as Parfit believes, then how could that mental action be described impersonally? Parfit does not say. But if the thinking of a person's thoughts is potentially observable by another person, then the person's life may truly be described in impersonal terms.

Parfit's self-reductionism and its implications occupy the middle of the book. They are preceded by extensive discussions of personal and moral dilemmas, time, and rationality, and followed by an exploration of

[5] It is not possible to specify exactly what the pattern is that defines love. In this sense the meaning of ''love'' is like the meaning of ''game.'' As Wittgenstein (1958) noted, a game is a family of activities that cannot be precisely defined. The same is true of love; but love is nonetheless a behavioral pattern.

what we owe to future generations. All of this ought to be of great interest to behaviorists. Working from the inside out, Parfit discusses personal dilemmas first and moral dilemmas second.

*Personal Dilemmas*

Personal dilemmas are what a psychologist would call self-control dilemmas. They involve attempts to follow the apparently simple Self-Interest Theory, S (p. 4): ''For each person there is one supremely rational ultimate aim: that his life go, for him, as well as possible.'' Consistent with S, a person might be a hedonist, aiming to maximize happiness; he might be interested in maximizing fulfillment of certain desires such as to be successful; or he might just have an objective list of high-valued states such as self-development, knowledge, awareness of true beauty, etc. S is thus not a pure pleasure theory. Dilemmas arise, however, when one sub-aim conflicts with another. Here is an example (p. 6):

> Kate is a Writer. Her strongest desire is that her books be as good as possible. Because she cares so much about the quality of her books, she finds her work rewarding. If her desire to write good books was much weaker, she would find her work boring. She knows this, and she accepts the Hedonistic Theory about self-interest [''What would be best for someone is what would give him the most happiness.'' (p. 4)] She therefore believes that it is better for her that her strongest desire is that her books be as good as possible. But, because of the strength of this desire, she often works so hard, and for so long, that she collapses with exhaustion, and is, for a period, very depressed.

We all know people like Kate. She is a *workaholic* (a word Parfit does not use). He adds (p. 14):

> It is best for Kate that her strongest desire is that her books be as good as possible. But, because this is true, she often works very hard, making herself, for a period, exhausted and depressed. Because Kate is a Hedonist, she believes that, when she acts in this way, she is doing what is worse for her. Because she also accepts S, Kate believes that, in these cases, she is acting irrationally….
> She can claim: 'When I do what I believe will be worse for me, my act is irrational. But, because I am acting on a set of motives that it would be irrational to cause myself to lose, I am *not* irrational. More precisely, I am *rationally irrational.*

I have ranked the two patterns of Kate's behavior below:

1. *Work very hard to write the very best books possible at the cost of frequent bouts of anxiety and depression.*
2. *Work significantly less hard to write less-good books at the cost of frequent periods of boredom.*

Kate values Pattern-1 over Pattern-2. However, we may assume that, in and of itself, she values working less hard over working very hard and that, in and of itself, boredom is less aversive than anxiety and depression. According to Parfit, although Kate *as a person* is not behaving irrationally when she chooses Pattern-1, her *specific acts* are irrational because, as individual acts, they are each less valuable than their alternatives. What weights the decision in favor of Pattern-1 is the higher quality of the books that result from that behavioral pattern. Parfit seems to be saying that book quality is somehow more objective and less personal than amount of work, anxiety, depression, or boredom. But Kate would be a highly unusual writer if what she valued were simply the quality of the resulting books. Most writers not only want their books to be good but also want to be the person who wrote the books and, usually, to be recognized as such. The example of Kate is less believable if we have to imagine that she would make the sacrifices she does in order that some stranger be enabled to write a better book. What does make a difference between writing a high-quality book and being bored, being depressed, or being anxious is that writing a high quality book is a more abstract and long-lasting action than the others. Given Kate's history, the hard work, anxiety, and depression are *components* of writing a high-quality book. She cannot have the one without the others.

Dilemmas like Kate's are very common in self-control situations. Figure 1 is a version of Herrnstein and Prelec's (1992) ''Primrose Path'' of addiction. It diagrams a highly simplified alcoholic's dilemma. As the alcoholic's drinking rate increases, he builds up tolerance to alcohol and the value of each individual drink decreases (line B–C). At the same time, as his drinking rate increases above

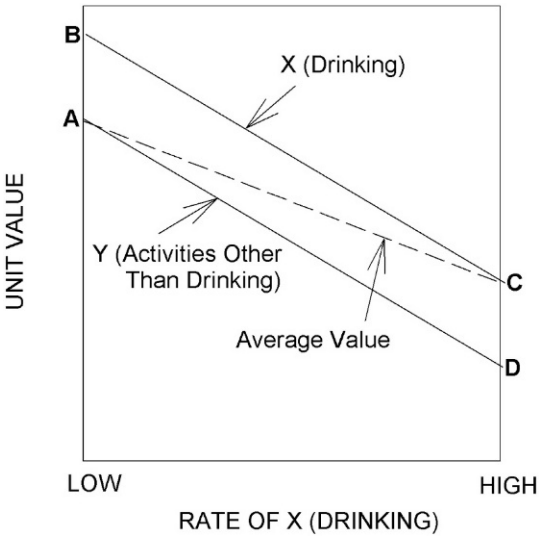


Fig. 1. Alcoholic's Dilemma. The value of an individual drink is always higher than that of not-drinking (solid lines) but over time not drinking is more valuable than drinking (dashed line).

a moderate level, his life gets worse and worse. His health, his social relations, his job performance, all decrease. The higher his rate of drinking, the more miserable he is at moments when he is not drinking (line A–D). At every point, regardless of the alcoholic's rate of drinking, having a drink is better than not having it. In order to reach or stay at Point A, the point of highest average value, the alcoholic must repeatedly choose the lower-valued of the available alternatives. This is his dilemma. In order to reach and stay at Point A, the alcoholic must be, in Parfit's terms, *rationally irrational*; in order to obtain a more abstract and high-valued state, his particular acts must be, according to Parfit, irrational. Drinking moderately is, for the alcoholic, like writing a high-quality book for Kate. It is a high-valued behavioral pattern composed of relatively low-valued particular acts. As Parfit points out, simple Self-Interest Theory (S) will not solve the dilemma. Such dilemmas are, according to Parfit, *self-defeating*. If you desire to base your behavior with regard to your future self on a set of moral principles, S is not sufficient. Moreover, since self-control and morality are twin processes, equivalently simple moral rules will also be found to be self-defeating.

*Moral Dilemmas*

Here is a simple moral rule corresponding, in the moral sphere, to S in the personal sphere. Parfit calls it Consequentialism (p. 24):

> (C1) There is one ultimate moral aim: that outcomes be as good as possible.
> (C2) What each of us ought to do is whatever would make the outcome best, and
> (C3) If someone does what he believes would make the outcome worse, he is acting wrongly.

This seems simple enough but, like S, it turns out to be self-defeating. For example, suppose a person is a pure Utilitarian; she accepts C1, C2, and C3, and she believes that the greatest good for the greatest number is the best of all possible outcomes. Clare is such a person (p. 32):

> Clare could either give her child some benefit, or give much greater benefits to some unfortunate stranger. Because she loves her child, she benefits him rather than the stranger.

Clare seems to be violating her own belief in pure Utilitarianism. The greatest good for the greatest number would seem to dictate the opposite choice. But it is a good thing for society for parents to be biased in favor of their own children. (Evolutionary considerations could provide justification for that belief but Parfit ignores such considerations and simply states as a fact Clare's preference for her child.) Although by her own standards her specific act is immoral, in the context of what is ultimately best for society her particular act is perfectly moral. According to Parfit, just as Kate's act was *rationally irrational*, Clare's act is *morally immoral*. For Parfit, the basic inconsistency is between a person's particular (irrational or immoral) *act* and a (rational or moral) *person*. For Parfit, the rationality and morality lie in Kate's and Clare's motives or dispositions (conceived as interior states). Their motives, along with their beliefs, are said to cause their acts. But, for teleological behaviorism, motives cannot cause acts; Clare's love for her child cannot be the *efficient cause* of her favoritism (but may be its *final cause*; Clare's love for her child is a pattern in her behavior; this particular act is part of that pattern).

Moral dilemmas are familiar to psychologists in the form of prisoner's dilemmas (Rachlin,

Brown & Baker, 2000). The prisoner's dilemma most commonly studied in the laboratory involves two players, each with two alternatives: *cooperate* or *defect*. If both players cooperate, both get moderately high rewards; if both defect, both get moderately low rewards; if one cooperates and the other defects, the cooperator gets a very low reward and the defector gets very high reward. (Corresponding contingencies may be programmed with punishers or mixed rewards and punishers.) The reward amounts are such that each player gains most (or loses least) by defecting, regardless of what the other player chooses. If two-person, prisoner's-dilemma games are played repeatedly, the usual outcome is that both players end up defecting whereas, if they had both cooperated, they would have gained more (or lost less). Although two-player games are typically studied in the laboratory, many-person dilemmas are much more common in everyday life, as Parfit points out. He gives some examples (pp. 61–62):

> *Commuters*: Each goes faster if he drives, but if all drive each goes slower than if all take busses;
> *Soldiers*: Each will be safer if he turns and runs, but if all do more will be killed than if none do;
> *Fishermen*: When the sea is overfished, it can be better for each if he tries to catch more, worse for each if all do;
> *Peasants*: When the land is overcrowded, it can be better for each if he or she has more children, worse for each if all do….
> There are countless other cases. It can be better for each if he adds to pollution, uses more energy, jumps queues, and breaks agreements; but if all do these things, that can be worse for each than if none do….In most of these cases the following is true. If each rather than none does what will be better for himself, *or his family, or those he loves*, this will be worse for everyone.

Consider the 10-person prisoner's dilemma game I play with audiences whenever I give a lecture or talk at a colloquium. First I say that the purpose of the exercise is to get everyone into the same mood—and that mood is ambivalence. Then I hand out blank index cards to 10 random members of the audience and I ask everybody else (as I ask the reader) to imagine that they had gotten a card. I then say that their job is to write either an *X* or a *Y* on the card, as they choose, subject to the following rules:

1. If you choose *Y* you will receive $100 times *Z*.
2. If you choose *X* you will receive $100 times *Z* plus a bonus of $300.
3. *Z* equals the number (of the 10 players) who choose *Y*.

I say, regretfully, that the money is purely hypothetical and then I point out several properties of this game. First, I strongly emphasize that for any particular player it is always better to choose *X*. By choosing *X* a player will subtract 1 from *Z* and thereby lose $100 but more than make up for that loss by the $300 bonus. The net gain for a particular player by choosing *X* is therefore $200. *"And this holds regardless of what anyone else chooses."* I say this slowly, loudly, and twice. I pause after saying it and, as far as possible, stare each member of the audience in the eye, effectively daring them to dispute the point; but no one ever does challenge it. It is perfectly true. I then emphasize the point further by saying that any lawyer would advise them to choose *X*.

Then I say: ''However, if everyone obeyed their lawyers and chose *X*, *Z* would equal zero and each person would earn just $300 whereas if everyone disobeyed their lawyers and chose *Y*, *Z* would equal 10 and each person would earn $1,000. This is the dilemma.''

I point out, truly, that there is no right or wrong answer. Then I tell them to mark their cards, letting no one see what they have written. I say that if anyone asks what they have chosen they should flip a mental coin and reply *X* or *Y* randomly. Their choice will be completely anonymous. No one will ever know what they have chosen or how much they have (hypothetically) earned. Then I collect the cards.

I have played this game dozens of times: with college students; with professors of psychology; with philosophers; with Italian socialist economists; with Japanese psychologists. The median response is about 50:50, 5 *Y*'s to 5 *X*'s plus or minus 2 either way. I then put up the diagram shown in Figure 2 and use it to look up the earnings of the *X* and *Y* choosers. For example, with 5 *Y*-choosers, $Z = 5$ and each *Y*-chooser gets $500 while each of the 5 *X*-choosers gets $800. In prisoner's-dilemma terms, an *X*-choice is a defection'' and a *Y*-choice is a ''cooperation.'' Although the

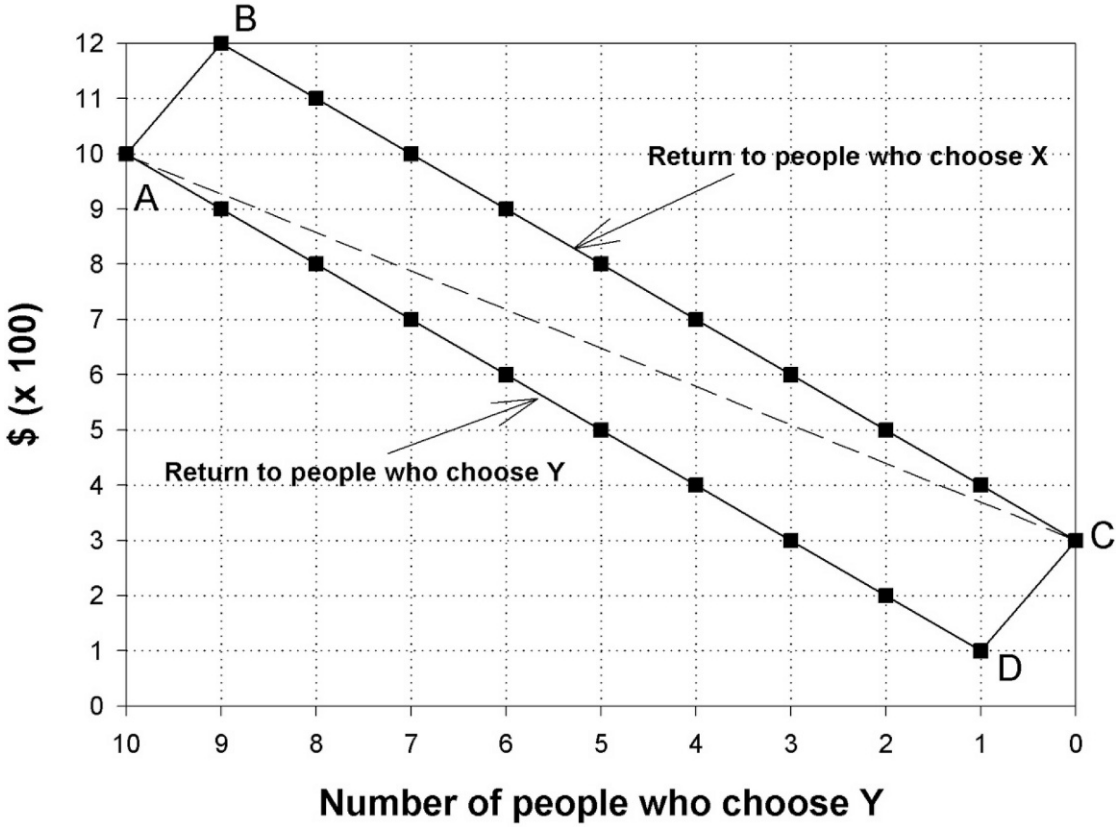


Fig. 2. Prisoner's Dilemma. For each individual player, Choice X earns more than Choice Y (solid lines). But, for the group as a whole, Choice Y earns more than Choice X (dashed line).

money earned is hypothetical, laboratory experiments with real money (albeit in lesser amounts) have found significant numbers of cooperations (*Y*-choices) in one-shot prisoner's dilemma games such as this one (Camerer, 2003).

Sometimes, depending on the circumstances, I try to lead a discussion (without anyone revealing what they actually chose) of why a person should choose *Y* in this game. They cannot claim that a *Y*-chooser's reputation would be enhanced, since the choices are completely anonymous. They cannot claim to be influencing other players' choices in future games, since they know that the game will be played only once during my lecture and then the audience will disperse. Many in the audience do say, despite all my efforts to forestall it, that they would choose *Y* if they believed that everyone else would choose *Y*. This rationale, on the face of it, is a *non sequitur*. Since the difference between the outcomes of an *X* and a *Y* choice is absolutely constant, what anyone else chooses in this game should, in theory, not influence their choice. What others do in this game will have a strong effect on a player's earnings. But what others do is irrelevant to what *you* should (in theory) do – you ''should'' defect. It is certainly incontrovertible that if a person chooses *Y* he will be happy if everyone else chooses *Y* and he earns $1,000. But would he be happier in such a case than he would be if he alone had chosen *X* and earned $1,200? If so, why? It seems as if the enhanced earnings of *others* (the extra $900 paid to the other players due to each player's *Y*-choice) at a sacrifice of $200 for themselves has an effect on people's behavior.

For coping with such dilemmas in everyday life, Parfit has an appealing suggestion (pp. 100-101). He suggests first determining what the outcome would be for you if everyone, including you, defected. In the case

of my lecture game, that would be the point where $Z = 0$ (Point C in Figure 2) resulting in $300 for you. Then you should determine how many other players would have to cooperate in order for you to obtain the same outcome ($300) if you also cooperated. In the game of Figure 2, if a total of 3 players chose $Y$ ($Z = 3$), each $Y$-chooser would earn $300. So, in order for you to earn $300 or more by cooperating, at least 2 *other* players would also have to cooperate. Parfit calls this number, $k$ but I will call it $h$ so as not to confuse it with a discounting constant ($k$) to be introduced later. In Figure 2, $h = 2$. Finally, you should estimate, as best you can, the number of other players who will actually cooperate. If that number is equal to or greater than $h$, you should cooperate too. However, as Parfit admits, his suggestion does not actually resolve the dilemma. If two other players chose $Y$ in my lecture game, and you also chose $Y$, you would earn $300 whereas, if you had chosen $X$, you would have earned $500.

By following Parfit's suggestion (and if your prediction is not too far off) you avoid cooperating while almost everyone else defects (being a sucker) or defecting while almost everyone else cooperates (being a free-loader). Why should you want to avoid either of these extremes, given that your choice is anonymous? You would have to have reasons extrinsic to the problem itself; for example, you may believe that the time and effort needed to determine whether every such choice you make in life will be truly and permanently anonymous is greater than the cost involved in following Parfit's suggestion (in my lecture game, $200 plus calculation costs). If this were the case, you should follow Parfit's suggestion as a general rule.[6]

Recall Clare's dilemma and her resolution of it. She believes that she violates strict Utilitarianism (the greatest good for the greatest number) by favoring her own child. However, her behavior (we assume) is consistent with a higher principle than strict Utilitarianism (whether or not she can verbalize it): Society will not function smoothly unless parents strongly favor their own children. But that principle is not all-or-none. Society also benefits if siblings favor each other and, to a lesser extent, if grandparents favor their grandchildren. In nonhuman societies, such biases have evolved so as to depend very strongly on genetic overlap. Genetic overlap is important in human societies as well, but physical proximity and common interest are also factors in determining the degree to which one person favors another (Rachlin & Jones, 2008a). In other words, there exists for each of us, a *social discount function* that characterizes our social choices.

Social discount functions measure the reduction in value to Person A of a reward given to Person B as a function of the social distance between them (Jones & Rachlin, 2006, 2009; Rachlin, 1989; Rachlin & Jones, 2008a, 2008b, 2009).[7] Such a bias is consistent with what Parfit calls *Common-Sense Morality* (M). In favoring her child, Clare puts Common-Sense Morality above Consequentialism, which is neutral with respect to social distance. Parfit says (p. 100):

> According to Common-Sense Morality, we ought to give some kinds of priority to the interests of those people to whom we are related in certain ways. Besides our children, some examples are: our parents, pupils, patients, clients, those whom we represent, and our fellow citizens. Let us say that we are *M-related* to these people.

Like S and C, M may be self-defeating in Parfit's terms. Consider a school fund (say for an outing) where the beneficiaries are the children of the potential contributors. If a few parents failed to contribute, the outing would take place anyway and their children would benefit along with the rest. Each set of parents thus faces a prisoner's dilemma, one step removed (for their children rather than for themselves). In such a case, according to Parfit (p. 102), "…each of us ought to give *no*

[6] Unfortunately, Parfit's suggestion will not work in self-control dilemmas such as the alcoholic's dilemma (Figure 1). If you are at Point C (an alcoholic currently drinking at a very high rate) there is no way to get directly to another point on the diagram without waiting for time to pass without drinking. To get back to Point A, for instance, you must follow the path C-D-A. This takes time and at every point along the way it will always be immediately better to drink than not to drink.

[7] Parfit, along with some political scientists and economists, discusses a measure called "The Social Discount Rate" (pp. 357, 480–486). The Social Discount Rate refers to *delay* discounting of social goods. This rate is conceptually different from *social discounting* as I use the term—the reduction in value to Person A of a reward given to Person B as a function of the *social* distance between them.

priority to those to whom he is M-related.'' That is, each parent ought to contribute to the fund (unless, following Parfit's suggestion, she believes that, even if she contributes, so few other parents will contribute that the outing will not take place).

Assuming that Common-Sense Morality implies the existence of a social discount function, people may have a reason to contribute in a multiperson prisoner's dilemma situation *independent* of whether or not anyone else contributes. In my laboratory at Stony Brook we have developed a method to obtain social discount functions. Participants are first asked to imagine that they rank-ordered the 100 people closest to them. This rank order is our measure of the social distance between the participant (Person A) and another person (Person B). Then, for selected social distances (usually ranks 1, 2, 5, 10, 20, 50, and 100) in turn, we ask the participant (Person A) to choose between a fixed amount of (usually hypothetical) money for the person at that social distance (Person B), and a varying amount of money for themselves. For example, we may ask, ''Which would you prefer, \$100 for the 10$^{th}$ person on your list or \$100 for yourself?'' Almost all participants, of course, prefer \$100 for themselves. On the other hand, almost all prefer the same \$100 for the 10$^{th}$ person on their list to \$0 for themselves. At some point between \$0 and \$100 for themselves, people must be indifferent between a (usually) lesser amount for themselves and \$100 for the 10$^{th}$ person on their list. We determine that point by varying, up or down, the lesser amount for themselves. We find that, as social distance increases, the crossover point decreases. In other words, the highest amount of money a person is willing to forgo for herself in order to give \$100 to another person decreases as a function of the social distance between them. A hyperbolic equation describes average data well ($R^2$ between 90% and 99% in different experiments):

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

where $V$ is the value of the fixed reward for Person B, $v$ is the crossover point—the value for Person A of $V$ given to Person B, $N$ is the social distance between Person A and Person B, and $k$ is a constant that differs from one person to another. The greater is $k$, the less is $v$, the more ''selfish'' is Person A—the less money the participant is willing to forgo for Person B's benefit. Typically, in our experiments, the fixed reward is set at \$75 but we also have used \$100 as in the example above. With either of these fixed rewards, the median $k$-value of the best fitting equation to individual-participant crossover points is about 0.05.

Equation 1 can now be used to calculate whether a person is likely to choose X or Y in my lecture game (Figure 2).[8] Suppose the game is being played among undergraduates taking introductory psychology. In independent questionnaires we have found that on average these students rate their social distance ($N$) to a random member of their class at about 75.[9] Recall that in my lecture game (Figure 2) each person's choice of $Y$ increases $Z$ by 1, giving \$100 to each of the 9 other players *regardless of their own choice*. Using Equation 1, we can calculate how much a $Y$-choice is worth to a subject. In Equation 1, with $V = \$100$, $k = 0.05$, and $N = 75$, $v$ may be calculated. With these numbers, $v = \$21$. That is, it is worth \$21 to an average introductory psychology student to give \$100 to one random classmate. Multiplying that amount by the nine other players, the average gross value of a $Y$-choice is \$189. This almost balances the \$200 cost of a $Y$-choice. Given the variance of the results of each stage of this calculation, it is fair to say that the costs and benefits of a $Y$-choice are about equal. Such equality would account for the approximate 50:50 split between $X$- and $Y$-choices in the dozens of times I have played this game in

[8] Equation 1 should not be thought of as the output of an internal mechanism that a person might be able to consult by introspection, still less as a template for conscious calculation. Rather, it is a way for an observer to describe the pattern of a person's choices among alternatives varying in social extent.

[9] This social distance was estimated by the following method: First as before we asked participants to imagine that they had made a list of the 100 people closest to them. Then we asked them to imagine all 100 people on their list standing with them on a large field at a (physical) distance from them proportional to their perceived social distance. We then determined the relationship between physical distance on the field ($F$, in feet) and rank order ($N$). [For average data, $F = 0.19N^{2.2}$ or $N = 3.3F^{0.45}$]. Then we asked participants to imagine a random member of their class on the field at a distance proportional to his or her social distance. On average that was about 2,500 feet, equivalent to an $N$ of about 75.

lectures. Players with *k*-values above 0.05 would be expected to choose *X* and players with *k*-values below 0.05 would be expected to choose *Y*. Over all, players' *k*-values should correlate negatively with their *Y*-choices in my lecture game. In a similar game (a public goods game) where players chose how much money to donate to a common fund, we found individual *k*-values to be inversely correlated with the amount of money participants were willing to contribute (Jones & Rachlin, 2009). As Parfit might say, it should be no more surprising that people are willing to make sacrifices (graduated by social proximity) for the benefit of other people than it is that people are willing to make sacrifices (graduated by temporal proximity) for their own future benefit.

With respect to *delay* discount functions, Parfit correctly implies that delay itself can have no effect on the subjective value of a reward. Certainly some physical process must occur during the delay period. Otherwise, time would have no meaning. If an entire person could be flash-frozen at the moment he put a dollar into a candy machine then flash-thawed (with no physical damage) at the moment of candy delivery, the delay between act and reward would be effectively zero no matter how long it actually was. For all we know, the entire universe could have been similarly flash-frozen and thawed just a moment ago—for a hundred ''years'' or a thousand. It would make no difference. Reward and punishment, apparently discounted by delay, may really be discounted by probability, by physical decay, by loss of use, by activities during the delay period, etc. Or by some innate brain process. The important fact is that delay discount functions have been obtained by methods similar to that described above for social discounting, that their functional form is hyperbolic (like Equation 1), that individuals differ in their degree of delay discounting (*k*) and, as you would expect, children, alcoholics, gamblers, smokers, heroin and cocaine addicts, and other groups who behave relatively impulsively in everyday life, have significantly higher individual *k*-values than adults, non-alcoholics, non-gamblers, etc. (See articles in Madden & Bickel, 2010). The task for society then would be not to convince people to obey some rule such as Parfit's suggestion but to develop methods to decrease the slope of their social and delay discount functions. It is true, as Parfit points out, that perfectly flat functions may lead to bad individual and social outcomes. But it seems safe to say that society is in much less danger from too-flat discount functions than from too-steep ones; the golden mean of discount functions is closer to the top than the bottom.

*Population Dilemmas*

There are population dilemmas that correspond in form to personal and moral dilemmas. Governments, for example, may encourage population growth or may attempt to keep the population constant, replacing deaths by births. Using a diagram similar to Figures 1 and 2 (p. 382), Parfit imagines a case where a growth policy always produces a higher average quality of life than replacement over the next two or three generations. Across several generations of repeated population growth, however, average quality of life deteriorates. Nevertheless, as it deteriorates, growth remains the better option over the next few generations; replacement does not pay off in higher quality of life unless it persists for several years. A growth policy, in Parfit's hypothetical example, is like drinking for an alcoholic or like an *X*-choice in my lecture game; a replacement policy is like refusing a drink or like a *Y*-choice.

Parfit argues that we should value the quality and quantity of life of generations in the distant future no less than the quality of life of generations in the near future. As with delay discounting, simple passage of time, he believes, can have no effect by itself. However, as with delay discounting, many factors other than time itself operate over time. The people 1,000 generations from now, for instance, will almost certainly have needs and desires so different from ours, that it would be foolish to give our guesses as to what those needs and desires might be as much weight as our knowledge of the needs of people two or three generations from now. What Parfit calls ''The Social Discount Rate'' (the *delay* discount rate of social goods), like the delay discount rate of individuals, should be low. In any case, even in Parfit's conception, quality of life of future generations would have to be discounted by the probability that our current estimate of its components is correct. And that probability must decrease over future generations.

Much of the final section of the book consists of an attempt to find a rule that will balance quality of life with the number (quantity) of people living in future generations. If only we could ignore quantity, and focus merely on average quality of life, the following rule would suffice (p. 369):

> Q: If in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have lower quality of life, than those who would have lived.

Problems arise, according to Parfit, if the number of people who would ever live is different for the two alternatives – if *quantity* must be considered along with quality. How should you weigh the relative importance of quality and quantity? In the face of this difficult problem, one is tempted to just ignore quantity and focus wholly on quality. Parfit rejects this solution; he says (p. 401–402):

> In one imaginary history, Eve and Adam have lives that are very well worth living, and have no children. They are the only people who ever live. In a different possible history, millions of people have lives that are very well worth living, but none have lives that are quite as good as those that Adam and Eve would have had. In this second possible history, the [average] quality of life is slightly lower, but there is a very much greater quantity of both happiness and of whatever else makes life worth living. On the…claim that only quality matters, the first possible history would be better. There are some people who accept this conclusion.

But Parfit is not one of them, and he assumes that the reader also does not accept it. But there is something odd (to me at least) about the argument's main premise. I cannot imagine being happy as Adam or Eve. For me, perfect happiness (my Eden) would require not only the country and/or the seashore but also a wide social life, including spouse, children and friends. It also requires museums, novels, movies, operas, a good hi-fi set, good restaurants, the web, Google, Netflix, and so on. These in turn require artists, musicians, physicists, mathematicians, engineers, etc. And, to make me perfectly happy, my world would not contain just any physicists, mathematicians, novelists, musicians, movie directors, chefs, but geniuses at their professions. But in any real world I can imagine (that is, in a Darwinian world) millions and millions of ordinary people would be required in order to produce sufficient numbers of geniuses. I am not being Dr. Pangloss here. I can certainly imagine a world better than the one we have—but not with just two people. Adam or Eve's happiness, understood as an internal state that could exist at a moment, seems to me not much more worth having than the happiness attainable by taking a drug or the happiness to which H.L Mencken referred:

> *Happiness is probably only a passing accident. For a moment or two the organism is irritated so little that it is not conscious of it; for the duration of that moment it is happy. Thus a hog is always happier than a man and a bacillus is happier than a hog* (quoted by Evans, 1968, p. 300.)

Further qualities that Parfit may consider to be part of a life worth living such as, ''self-development, knowledge, awareness of true beauty,'' can be no more complex than society allows them to be. A chimpanzee (if not a hog), living in a community of other chimpanzees, would have a better chance of attaining them than would Adam and Eve running around naked in the jungle with only each other for company. Even understood as temporally extended patterns of overt behavior, these qualities, in a society of two people, would be primitive versions of our conceptions of them. Contrary to Parfit, I think we can ignore quantity as a separate factor in human happiness—because quality implies quantity.

### *Toward a Behavioral Ethics*

It is not possible, of course, to construct a behavioral ethics in the space of this review, but it is possible to consider how such an ethics might develop. I will follow Parfit's lead by first considering individual behavior over time—self-control dilemmas—where the issues are perhaps clearer, and afterwards consider the much more difficult case of behavior over social space—moral dilemmas.

*Self-control.* Imagine a person, Jane, in the following situations. In the first situation Jane is always free to choose repeatedly between behavior X and behavior Y (having an alcoholic drink and not having one, for example). After each choice she may immediately choose again between the same two alternatives. Over a period of time (say a month) she will exhibit

a pattern that we simplify as the rate, *X/(X+Y)*. Now imagine that each morning she must choose the pattern, *X/(X+Y)*, for that day. She may choose as much or more X's and Y's as she did before but, over a period of a day, she must make all choices in advance and she is committed for the day to her initially chosen pattern. Now extend the period of commitment to a week and finally to the whole month. To the extent that her pattern over the month with free choices, or with a narrower commitment, matches the pattern she chose with a more extended commitment, we may say that Jane's choices with the narrower commitment were self-controlled —she was doing what she intended to do, what she values most over the longer term. Such a test would measure Jane's self-control directly; but in real-life situations it would be impractical if not impossible to administer. Behaviorists would have to develop other tests—such as delay discount functions—that approximated results of the proposed direct measure of Jane's self-control.[10]

For teleological behaviorism, individual acts are neither self-controlled nor impulsive in themselves; the nature of individual acts is determined by the pattern in which they are imbedded. Since this pattern extends into the future, the nature of an individual act is not completely determined (neither by an observer nor by the actor herself) when it is emitted. For Parfit, on the other hand, the nature of an individual act, caused as it is by an internal prior event, must already be determined even before it is emitted. This predetermination in Parfit's theory of self-control is, I believe, what causes logical problems and prevents him from developing a rational approach to self-control.

As a more realistic problem of self-control than Jane's, consider a dilemma similar to that of Parfit's Kate (discussed previously). John is a recovering alcoholic with a history of vacillation among three behavioral patterns:

1. Drunkenness. *John consumes 10 to 15 alcoholic drinks a day—or sometimes more; his personal relations, his health, and his work all deteriorate.*
2. Teetotaling. *John swears off drinking entirely—he becomes a teetotaler. His health and his work recover over time but his personal relations are not as good as they might be. At parties or when going out for dinner with friends he is a kind of wet blanket; his friendships are not as close as in the past when he drank moderately. This pattern is much better for him than is Pattern 1 but not as good as Pattern 3, social drinking.*
3. Social drinking. *John "lets himself go" a little at occasional parties, and drinks a couple of glasses of wine when eating out with friends but at home confines himself to a single glass of wine with dinner.*

In the very short term, Pattern-1 is the most valuable. John prefers drinking now to not drinking now. But, if required to commit to a pattern for the next 5 years, John would choose Pattern 2 over Pattern 1 and would choose Pattern 3 over either of the other two. Pattern 3 is the most difficult to maintain because, within it, John has to switch many times from a high-valued, short-term activity (drinking) to a lower-valued, short-term activity (not drinking). Parfit would say that John, when exhibiting Pattern 3, and drinking at a party, is being (like Kate when she works hard) *rationally irrational*. For a teleological behaviorist, however, there is nothing irrational or lacking in self-control in John's behavior during Pattern 3 (or Kate's behavior when she works hard).

Parfit would see John as (at least partially) irrational during Pattern 3 because, to repeat, for Parfit, value is a property of an internal and current state. Pains are bad and pleasures are good, according to Parfit, only to the extent that they are actually experienced; value over the long term is, for him, the sum of the values of momentary (internal) states. Those momentary states may be highly complex and abstract (for instance, desire for self-development, knowledge, awareness of true beauty). Nevertheless, a person's mental state over the long term, for Parfit, must be no more than the sum of a series of immediate states (however complex) and its value the sum of their values. A person who deliberately chooses a long-term pattern consisting of low-valued immediate states in preference to one consist-

[10] This behavioral test would measure what Jane does relative to what she wants while Parfit is interested in what she does relative to what it is good for her to want. The question of what it is good for a person to want, however, seems to me to go beyond philosophy as well as psychology and into issues of medicine and hygiene.

ing of higher-valued immediate states is, therefore, according to Parfit, not completely rational. For a teleological behaviorist, on the other hand, it is not only possible for a long-duration pattern to contain lower-valued components, a more valuable long-duration pattern may consist *solely* of lower-valued components (as illustrated in Figure 1).

This difference between Parfit and teleological behaviorism is fundamental (Rachlin, 1994). For Parfit as well as most other contemporary philosophers and nonbehavioral psychologists, discrete acts may be explained in terms of the internal and equally discrete mental state that causes them (their efficient cause); for teleological behaviorists, discrete acts may be explained in terms of the temporally extended pattern of overt acts in which they are imbedded (their final cause). A person who makes decisions consistent with valuable long-term patterns is said by teleological behaviorism to be self-controlled; the longer the pattern, the more self-controlled. Self-control, thus understood, is a relative term. No act or pattern of acts can be self-controlled in isolation, but only relative to other acts or patterns. John's teetotalism maintained over 5 years would therefore be more self-controlled than drunkenness over that period but less self-controlled than social drinking.

In practice it is not possible to offer a person (Jane in our example above) choices under various degrees of constraint. How then may we determine her degree of self-control? In theory, the temporal extent of the pattern of choices over which a person maximizes value may be measured by the constants in that person's delay discount functions. We are still far from being able to use delay discount functions in this way but recent research showing that addicts of various kinds discount money more steeply than do nonaddicts is encouraging (Madden & Bickel, 2010). It is important to remember that, from a behavioral viewpoint, a delay discount function is a measure of the duration of a pattern and not a representation of an internal state that a person could ''use'' or refer to when making choices (as a bank might do when calculating interest). Thus, research on discount functions can lead to measurement of self-control. The issues needing to be resolved are empirical not theoretical. We know that alcoholics have steeper delay discount functions than do teetotalers and social drinkers. But do teetotalers have steeper discount functions than social drinkers? How do delay discount functions of different commodities relate to one another? How malleable are a person's delay discount functions? How are they changed by simple and complex verbal rules (discriminative stimuli)? How do delay discount functions relate to social discount functions? How does delay of a social reward relate to delay of a monetary reward? All these are empirical questions—answering them will constitute the development of a behavioral theory of self-control.

How *should* people behave to achieve ideal self-control? A behavioral theory might say, with perfect consistency, that people should behave so as to maximize value over the longest possible duration. The purpose of discount functions in such a theory would be to measure the duration over which value is maximized—the degree to which self-control approaches the ideal. It is that duration, not the slope of the discount function, that *defines* self-control. With this in mind, there are serious problems with discount functions as they are now obtained. It would seem that, the flatter the function, the greater the duration of the behavioral pattern, the more self-controlled the behavior. This would imply that completely flat functions would be indicative of ideally self-controlled behavior. Yet, as discount functions are now obtained, a person with a completely flat function would prefer $1,001 in 10 years to $1,000 today. Such a choice is clearly irrational. But this only means that we need to develop better measures of pattern duration, including probability, interest, forgone activities, etc. Parfit might argue that if we did include all of these factors, we would not need to include time itself. Perhaps, but remember it is the *duration* of the pattern over which value is maximized that we are attempting to measure.

*Morality.* As I have been arguing, in agreement with Parfit, moral dilemmas are similar in structure to self-control dilemmas. But moral dilemmas are more complex. It was relatively easy to devise a (theoretically) simple method to decide what Jane, in our example, wanted in the long run. An equivalent method for morality would have to address the more difficult question, ''What does a group of people want in the long run?''

Morality and immorality, like self-control and impulsiveness, are properties of patterns, not acts. In the flight attendant's spiel before takeoff she says that, when the oxygen masks drop down, parents should put on their own masks before putting them on their children. Otherwise a parent might lose the presence of mind to properly fix the child's mask. In an actual emergency, a parent who cares more for herself than for her child would behave identically to one who cares more for her child than for herself. To gauge a parent's degree of morality we would need a larger sample of his behavior.

> Recall Clare's dilemma as set forth by Parfit (p. 32):
> Clare could either give her child some benefit, or give much greater benefits to some unfortunate stranger. Because she loves her child, she benefits him rather than the stranger.

Clare's behavior is seen by Parfit as inconsistent with her ideals – she is *morally immoral*. But, depending on the wider pattern of her behavior, teleological behaviorism may see her behavior as completely moral. If the alternatives she chooses are generally consistent with what is on the whole good for society, and if parents showing preference for their own children is generally good for society, then this particular instance of that preference is in no way immoral.

Clare's preference for her child is like John's drinking at parties. Whether Clare's particular act is selfish or not depends on Clare's other behavior and not on her internal state; whether John's particular drink is impulsive or self-controlled depends wholly on his other behavior (whether this drink is part of a pattern of drunkenness or part of a pattern of social drinking) and not at all on his internal state while he is having the drink.[11]

Clare's behavior illustrates the complexity of moral dilemmas relative to self-control dilemmas. Her preference for her child is consistent with a wide social concern—what is good for society, as well as a narrow social concern—what is good for her immediate family. As indicated above, whether this particular act is actually part of a pattern of wide or narrow social responsibility depends on her other acts spread out over time, past and future. The morality of a pattern of acts depends not just on the *social* extent of that pattern but also on its *temporal* extent.

Whereas self-control is potentially measurable by the shallowness of two-dimensional delay discount functions, measurement of morality would require a three-dimensional surface. A behavioral theory of morality would parallel, in the sociotemporal sphere, a behavioral theory of self-control in the temporal sphere. Just as patterns that maximize value over greater temporal spans are more self-controlled than patterns that maximize value over lesser temporal spans, so patterns that maximize value over greater sociotemporal extents are more moral than patterns that maximize value over lesser sociotemporal extents. Just as delay discount functions may eventually be used to measure individuals' temporal spans, so sociotemporal discount functions may eventually be used to measure individuals' social spans. To my knowledge no empirical study of sociotemporal discounting has been published but, as described in a previous section, a method of obtaining individual social discount functions has recently been developed (Jones & Rachlin, 2006, 2009; Rachlin & Jones, 2008a, 2008b, 2009). Problems with this method parallel those with delay discount functions and are compounded by the additional dimension. But, as with delay discounting, the work of development is largely empirical. Many individual actions may be part of moral patterns or immoral patterns (Clare's favoritism for her child for example). The only way to decide for sure between one interpretation and another is to observe Clare's actions over a longer time period. How extended is the sociotemporal space over which her actions maximize value? That is the moral bottom line.

But those situations may be difficult or impossible to observe. Eventually we may develop a behavioral measure—a veridical sociotemporal discount function—to aid in determining an individual's socio-temporal space. I would guess that, if and when such a measure is developed, the data required to

[11] Of course both John's and Clare's specific acts are efficiently caused by immediate internal events. But, for teleological behaviorism, these internal causes are physiological events—actions of the nervous and muscular systems—not mental events. Mental events are patterns of overt behavior. To the extent that mental events may be said to cause behavior, they are final, not efficient causes (Rachlin, 1994).

make such determinations will be the behavior (verbal and nonverbal) of other people (friends and relatives) as well as that of the person herself. But again this is an empirical rather than a theoretical issue.

Meanwhile, on what basis may we behaviorists make moral judgments about our own and others' lives? Marvin Frankel and I (Frankel & Rachlin, in press) suggest evaluating the social and temporal breadth of the ''metadiscriminative stimuli'' that control our behavior—the social contingencies, social laws and moral codes that govern our lives—and our consistency in behaving in accordance with them. Moral behavior (what we call ''moral achievement'') consists in integrating the various stimulus-defined situations that control different behavior patterns over our lifetimes. We cite examples of people who act one way under one set of circumstances and another way under another set of circumstances without integrating the two patterns into a larger coherent framework. One example is Billy/William, an interesting, exciting, social risk-taker (''Billy'') in the presence of his old friend (MF) and a conservative conformist (''William'') in the presence of his wife. Billy/William suffered a (mild and temporary) nervous breakdown after a call from his wife interrupted an intimate conversation with his old friend. Another example is the college student at graduation when her teacher (who treats her as a responsible adult) meets her parents (who treat her as an irresponsible child). A far more serious example is the eponymous Herr Stauffen, a mild postmaster in prewar Germany who becomes a vicious Gestapo officer during the Nazi era. From our viewpoint, the observed incoherence is a sign of failure to develop (consciously or not) a moral framework (a metadiscriminative stimulus) that could guide behavior across situations:

> For the student, a childish display at graduation in front of a valued teacher or seemingly cold conduct before her beloved parents could put either of these relationships at risk. The consequences, in the case of Billy, may be the loss of his wife or friend. In the case of Herr Stauffen, the consequences could be loss of his life.

Nevertheless, Parfit's larger point is well taken. There is no simple formula for individual good, for social good, or for the good of future generations. Calculations have to be made and balances struck. There are no easy answers to the questions this book raises. Parfit has only started down the road leading from reasons and persons to moral behavior. I believe that he puts difficulties in his own way. His reasons are impersonal but not impersonal enough. He is half a behaviorist, but that is not behavioristic enough. Behaviorists too have set off on this road (e.g., Baum, 2005, and most notably Skinner, 1948, 1953, 1957, 1971, 1974) but they did not pursue the logical intricacies in their way. This book provides transits and levels of logic to map the path. However, actual progress will require the heavy machinery of empirical study and behavior analysis.

## REFERENCES


Baum, W. (2005). *Understanding Behaviorism: Behavior, Culture, and Evolution,* 2nd ed. Malden, MA: Blackwell Publishing.

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction.* New York: Russell Sage Foundation.

Evans, B. (1968). *Dictionary of quotations.* New York: Delacorte Press.

Frankel, M., & Rachlin, H. (in press). Shaping the coherent self: A moral achievement. *Beliefs and Values: Understanding the Global Implications of Human Nature.*

Herrnstein, R. J., & Prelec, D. (1992). A theory of addiction. In G. Loewenstein, & J. Elster (Eds.). *Choice over time.* New York: Russell Sage Foundation.

Jones, B. A., & Rachlin, H. (2006). Social discounting. *Psychological Science, 17,* 283–286.

Jones, B. A., & Rachlin, H. (2009). Delay, probability, and social discounting in a public goods game. *Journal of the Experimental Analysis of Behavior, 91,* 61–74.

Madden, G. J., & Bickel, W. K. (Eds.) (2010). *Impulsivity: The behavioral and neurological science of discounting.* Washington D.C.: American Psychological Association.

Parfit, D. (1984). *Reasons and persons.* Oxford: Oxford University Press.

Rachlin, H. (1985). Pain and behavior. *The Behavioral and Brain Sciences, 8,* 43–52.

Rachlin, H. (1989). *Judgment, Decision and Choice* W. H. Freeman & Co.: New York,

Rachlin, H. (1994). *Behavior and mind: The roots of modern psychology.* New York: Oxford University Press.

Rachlin, H. (2002). Altruism and selfishness. *The Behavioral and Brain Sciences, 25,* 239–296.

Rachlin, H., Brown, J., & Baker, F. (2000). Reinforcement and punishment in the prisoner's dilemma. In D. Medin (Ed.). *The psychology of learning and motivation: advances in research and theory, 40* (pp. 327–364). New York: Academic Press.

Rachlin, H., & Frankel, M. (2009). Taking pragmatism seriously: A review of William Baum's Understanding Behaviorism: Behavior, Culture, and Evolution (Second Edition). *Journal of the Experimental Analysis of Behavior, 92,* 131–137.

Rachlin, H., & Jones, B. A. (2008a). Altruism among - relatives and non-relatives. *Behavioural Processes*, *79*, 120–123.

Rachlin, H., & Jones, B. A. (2008b). Social discounting and delay discounting. *Behavioral Decision Making*, *21*, 29–43.

Rachlin, H., & Jones, B. A. (2009). The extended self. In G. J. Madden, & W. K. Bickel (Eds.). *Impulsivity: The Behavioral and Neurological Science of Discounting* (pp. 411–432). Washington DC: APA Books.

Skinner, B. F. (1948). *Walden two*. New York: Macmillan.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Skinner, B. F. (1971). *Beyond freedom and dignity*. New York: Knopf.

Skinner, B. F. (1974). *About behaviorism*. New York: Knopf.

Staddon, J. E. R. (2001). *The new behaviorism: Mind, mechanism, and society*. Philadelphia: Taylor & Francis.

Wittgenstein, L. (1958). Philosophical Investigations. G.E.M. Anscombe (Tr.) New York: Macmillan.